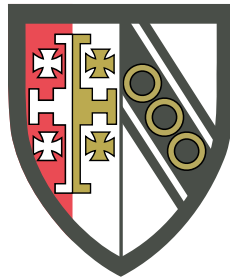




Endotype Discovery in Acute Respiratory Distress Syndrome



Romit Joy Samanta

Department of Medicine

This dissertation is submitted for the degree of
Doctor of Philosophy

Selwyn College

March 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Romit Joy Samanta
March 2021

Abstract

Endotype Discovery in Acute Respiratory Distress Syndrome

Dr Romit Samanta

Acute respiratory distress syndrome (ARDS) affects 10% of critical care patients and is characterised by acute refractory hypoxaemia and bilateral pulmonary infiltrates on thoracic imaging. Mortality from severe ARDS is approximately 40%, and has not changed in 50 years despite decades of study. Randomised controlled trials of therapies for ARDS have been unsuccessful due to the heterogeneity of the patient population. This has led repeatedly to potentially promising therapies being discarded. The primary reason for the failure is that the underlying biological processes occurring in ARDS are poorly understood.

This thesis attempts to address this heterogeneity, and explores the underlying biology by using an integrated, unsupervised bioinformatics approach to describe different mechanistic subtypes (endotypes) of ARDS. The endotypes described here are derived from analysis of data collected by three UK-based studies: an observational study of sepsis (GAinS), an observational study of severe influenza (MOSAIC), and a randomised controlled trial of simvastatin in ARDS (HARP-2).

A combination of automated clustering methods and network analysis tools have been used to integrate blood biomarkers and gene expression (transcriptomic) data to define distinct endotypes of ARDS.

Three endotypes of ARDS were identified in each of the studies. Integration of protein biomarker and transcriptomic data from patients recruited to the GAinS study identified three endotypes, one of which was characterised by severely dysregulated cytokine release, which we termed hyper-inflammatory. Two gene modules discriminated these patients from a hypo-inflammatory endotype, consisting of patients with globally depressed cytokine concentrations. Enrichment of the genes in these two modules identified genes that were important in vesicle fusion and cytokine release. Mutations in these genes cause the familial

type of haemophagocytic lymphohistiocytosis (HLH). The implication here is that these genes played a role in the severely dysregulated cytokine concentrations observed within patients with hyper-inflammatory, sepsis-associated ARDS.

Analysis of the cytokines and transcriptomic data collected during the MOSAIC study identified three endotypes we named: adaptive, endothelial leak and neutrophil driven. The endothelial leak endotype was characterised by enrichment of genes associated with SLIT-ROBO signalling. SLIT-ROBO signalling is essential for maintaining pulmonary endothelial integrity and failure of this mechanism has been shown to cause alveolar oedema in murine models of sepsis and influenza infection. These patients had significantly lower albumin levels than the adaptive endotype, and 48.5% of them required mechanical ventilation. Despite the greater need for mechanical ventilation, the outcomes of these patients were similar to those of patients with the adaptive endotype, of whom only 20.4% required mechanical ventilation.

Cluster analysis of patient biomarker concentrations from the HARP-2 study identified three endotypes. Two of these endotypes had elevated serum IL-6 and sTNFR-1 concentrations, consistent with a hyper-inflammatory profile. Patients with one of the hyper-inflammatory endotypes, which we termed MMP-8 dominant, demonstrated a strong therapeutic response to simvastatin compared with placebo (28-day survival, adjusted HR = 0.35, 95% CI 0.18-0.71; $p = 0.003$). Patients with this endotype, who received simvastatin, had a similar 28-day survival profile to patients with a hypo-inflammatory endotype, characterised by globally depressed biomarker levels. Patients with the other hyper-inflammatory endotype, which we termed sRAGE dominant, did not show any therapeutic response to simvastatin.

The endotypes described are temporally stable, and some relate to novel mechanisms not previously recognised in patients with ARDS. The endotypes are all biologically plausible, amenable to the development of further mechanistic insights using laboratory-based techniques, and may influence patient outcomes and response to treatments. Further development and prospective validation of these endotypes are required. If validated, they may offer the opportunity to stratify patients in future clinical trials to treatments that are more likely to improve their outcomes, whilst avoiding treatments that might cause adverse effects. The methods described in this thesis could be applied to other heterogeneous and poorly understood clinical syndromes.

I would like to dedicate this thesis to my loving parents Mita and Sukumar Samanta, and my wonderful, ever patient, always understanding family: Liz, Idris and Maya.

Acknowledgements

None of this work would have been possible without the support of so many people.

Dr Charlotte Summers, my supervisor, has provided me with incredible opportunities to explore my interests in critical care research, data science and systems biology for which I am eternally grateful. Building and maintaining the collaborations required to conduct this kind of research is challenging and demanding. Dr Summers has never made it apparent that the work she has been doing in the background to support this project was ever too much effort or a burden. I had not previously considered a career in research but with her guidance, I have undertaken this project with joy and no regrets. She has been the driving force behind the EMINENT:ARDS project and the results presented here are the product of her efforts.

Dr Adam Taylor from GSK, Stevenage has always been on hand to assist with bioinformatics support and ideas when I was stuck, early on in the project. Adam always had a suggestion for a new approach to hand and I am grateful for his advice and experience.

Professor Edwin Chilvers supported the EMINENT:ARDS project from its outset and before moving on from Cambridge provided valuable guidance and support to me.

I am grateful for the generous provision of resources from:

Professor Peter Openshaw, Imperial College London (MOSAIC study).

Professor Danny McAuley, Queen's University Belfast (HARP-2 trial)

Professor Julian Knight and Professor Charles Hinds, University of Oxford and Queen Mary University of London (GAinS study).

I would also like to acknowledge other members of the Knight group, University of Oxford: Dr Cyndi Goh, who helped me decipher the microarray data and Yuxin Mi who conducted the cytokine assays.

Other members of the EMINENT:ARDS consortium have always been supportive and positive in all my interactions with them, despite their busy schedules. Dr Claire Allen, Dr David Budd and Dr Andy Bayliffe from GSK have always made the resources of GSK available to me and provided guidance and support when requested. Professor Rachel Chambers, University College London, has been a major supporter of the EMINENT:ARDS project and helped me to draft manuscripts.

In the lab I have found friendship, discerning criticism of my nerdiness, and many enjoyable exchanges of ideas related and unrelated to science. Eléonore Fox and Rowena Jones both endured many hours of being stuck in an airless office with me whilst I struggled with computer code. They helped to support me when my programs failed with questionable music playlists, delicious baking and impromptu French lessons. Quinn Nelson was an exceptional masters student who learned fast and shared my passion for caffeine. Other members of the lab who were always on hand for valuable discussions include Dr Julian Subburayalu, Dr Neda Farahi, Dr Tony Ng, Dr Alistair Jubb and Dr Andrew Conway-Morris.

Rosalind Simmons, Andrew Savage and Katerina Stroud have always been obliging in response to my requests for their help.

Finally, I must acknowledge the patience and inspiration that my family have shown. Idris regularly inspected my graphs and asked questions that led to new insights, whilst Maya has always understood what ‘Daddy is doing work’ meant. My wife, Liz, has always been understanding, supportive and this work could not have been completed without her patience and love.

Funding

Funding for this PhD programme was from the Experimental Medicine Initiative to Explore New Therapies (EMINENT) consortium which is a collaboration between the Medical Research Council (MRC), GlaxoSmithKline PLC (GSK), the National Institute for Healthcare Research (NIHR) and six academic centres in the UK (Cambridge, Imperial College, Glasgow, Newcastle, Oxford and University College London).

The NIHR Cambridge Biomedical Research Centre, GSK, NIHR and MRC all contributed directly to support my salary and research costs for this PhD programme.

The views expressed in this work are those of the author and not necessarily those of the NIHR, GSK, MRC, University of Cambridge, NHS or Department of Health and Social Care.

Table of contents

List of figures	xv
List of tables	xix
Abbreviations	xxi
1 Introduction	1
1.1 Definition and features of acute respiratory distress syndrome	1
1.2 Epidemiology and patient outcomes	2
1.3 Pathophysiology	3
1.4 Treatment of ARDS	5
1.5 Stratification approaches in ARDS	8
1.5.1 Biomarkers in ARDS	8
1.5.2 Genomic studies in ARDS	12
1.5.3 Clinical variable-based subtypes of ARDS	17
1.6 Successful endotyping approaches in other diseases	20
1.6.1 Asthma	20
1.6.2 Breast cancer	21
1.6.3 Sepsis	22
1.7 Successful endotyping approaches in ARDS	25
1.8 Background to studies used in this thesis	27
1.9 Summary and aims	29
1.10 Hypothesis and thesis overview	30
2 Methods	31
2.1 Overview of methods	31
2.1.1 Data sources	34
2.1.2 Sampling times and patient status	36

2.1.3	Biological sample collection, processing and analysis	39
2.1.4	ARDS diagnosis	39
2.2	Hierarchical clustering	41
2.2.1	Determination of the optimum number of clusters	43
2.2.2	Management of missing data	46
2.2.3	Cluster stability	47
2.3	Microarrays	47
2.3.1	Preprocessing and management of batch effects	48
2.4	Weighted gene co-expression network analysis	50
2.4.1	Network construction and intuition	50
2.4.2	Downstream analysis of results after applying WGCNA	54
2.5	Data integration and linear discriminant analysis	55
2.5.1	Linear discriminant analysis	56
2.5.2	Assessment of LDA model performance	59
2.6	Endotype characterisation	60
2.6.1	Enrichment of gene lists	60
2.6.2	Determination of differential gene expression	61
2.7	General statistical methods	62
3	Clustering of biological data	65
3.1	Overview of results	65
3.2	Hierarchical clustering of protein biomarkers	66
3.2.1	Preprocessing and imputation	66
3.2.2	Hierarchical clustering: linkage methods	68
3.2.3	Hierarchical clustering of protein biomarker profiles from patients recruited to the GAINs study identified three clusters	72
3.2.4	Hierarchical clustering of protein biomarker profiles from patients recruited to the MOSAIC study identified three clusters	76
3.2.5	Hierarchical clustering of biomarker profiles in the HARP-2 patients did not identify an optimal number of clusters	80
3.2.6	Assessment of cluster stability	84
3.3	Discussion of protein biomarker clustering	84
3.4	Analysis of microarray data	88
3.4.1	Quality control and pre-processing	88
3.4.2	Differential gene expression analysis provides few insights in pa- tients with sepsis and ARDS	92
3.5	Weighted gene co-expression network analysis of microarray results	94

3.5.1	WGCNA identifies gene modules in the microarray results from the GAINs and MOSAIC studies	94
3.5.2	Module adjacency identifies closely related modules	101
3.5.3	There is no significant correlation between gene modules and traits in patients recruited to the GAINs study	102
3.5.4	Correlation between clinical variables and gene modules identify plausible biological processes in patients recruited to the MOSAIC study	104
3.6	Discussion of microarray and WGCNA results	108
4	Integration of protein biomarkers with transcriptomics	111
4.1	Differential gene expression between clusters	111
4.1.1	Differential gene expression between protein biomarker-based clusters of patients with ARDS in the GAINs study	111
4.1.2	Differential gene expression between protein biomarker-based clusters of patients with severe respiratory failure in the MOSAIC study	115
4.2	Gene module correlation analysis	121
4.2.1	There are no gene modules identified in the GAINs study that significantly correlated with protein biomarker clusters	121
4.2.2	Correlation between protein biomarkers clusters and gene modules identified important mechanisms in the 'red' MOSAIC cluster	122
4.3	Linear discriminant analysis of integrated biological data	127
4.3.1	Combining module eigengene and protein biomarker values preserves the properties of clusters	127
4.3.2	Linear discriminant analysis of ARDS samples between different clusters from the GAINs study	130
4.3.3	Neutrophil activation is an important discriminator of 'purple' and 'green' GAINs clusters in ARDS samples	133
4.3.4	Ranked discriminators of the 'yellow' and 'green' GAINs clusters in ARDS samples involve transcripts with important roles in immune function	135
4.3.5	Ranked discriminators of the 'yellow' and 'purple' GAINs clusters in ARDS samples do not identify a plausible mechanism to account for their differences	137
4.3.6	Further enrichment of key modules that discriminated the GAINs 'purple' and 'green' clusters identifies important sub-networks	139

4.3.7	Linear discriminant analysis of samples between different clusters from the MOSAIC study	145
4.3.8	Neutrophil degranulation discriminates the ‘grey’ and ‘red’ MOSAIC clusters	147
4.3.9	Regulation of SLITs and ROBOs discriminates the ‘red’ and ‘blue’ clusters from the MOSAIC study	149
4.3.10	Regulation of expression of SLITs and ROBOs discriminates the ‘blue’ and ‘grey’ MOSAIC clusters	153
4.3.11	Linear discriminant analysis of HARP-2 clusters	155
4.4	Summary and discussion: integration of biological data	157
4.4.1	Differential gene expression	157
4.4.2	Correlation between WGCNA-derived gene modules and clusters	158
4.4.3	LDA of clusters	159
5	Endotype characterisation	163
5.1	GAinS endotypes	163
5.1.1	With the exception of PaO ₂ -FiO ₂ ratio, there were no significant differences in organ dysfunction between the endotypes in the GAinS study	163
5.1.2	Patients in the GAinS study with the hypo-inflammatory endotype were more likely to have received steroid therapy	164
5.1.3	Survival analysis showed no significant differences between patients with different endotypes in the GAinS study	169
5.1.4	Endotypes identified in the GAinS study were not stable over measured time points	171
5.2	MOSAIC endotypes	172
5.2.1	Patients with the neutrophil driven endotype from the MOSAIC study had significantly worse multi-organ dysfunction.	172
5.2.2	Low albumin concentration is associated with the SLIT-ROBO endotype	176
5.2.3	MOSAIC endotypes are associated with different patient outcomes	176
5.2.4	MOSAIC endotypes were stable after 48 hours	180
5.2.5	The role of secondary bacterial infection in MOSAIC endotypes is uncertain	182
5.3	HARP-2 endotypes	184
5.3.1	Clinical features of endotypes identified in the HARP-2 study	184

5.3.2	HARP-2 endotypes are associated with different outcomes and treatment response	187
5.4	Summary and discussion of endotype characterisation	191
5.4.1	GAinS	191
5.4.2	MOSAIC	192
5.4.3	HARP-2	193
6	General Discussion and Conclusions	195
6.1	Summary of endotypes	195
6.2	Limitations	199
6.3	Validity of this approach in future studies	202
6.4	Summary and conclusion	203
	References	205
	Appendices	233
A	Ethical approvals	233
B	Details of microarray experiments	233
C	Methods for protein biomarker quantifications assays	234
D	Boxplots of protein biomarker concentrations in each cluster	237
D.1	GAinS	237
D.2	MOSAIC	240
D.3	HARP-2	244
E	Fully labelled biplots	245
F	Ranked linear discriminators between clusters from the GAinS study	246
G	Ranked linear discriminators between clusters from the MOSAIC study . .	250
H	Permissions	254

List of figures

1.1	Figure from Bos et al. (2019) showing gene expression PCA of patients with sepsis and ARDS subtypes	19
2.1	Flow chart of data analysis plan	33
2.2	Acute illness trajectory	38
2.3	Distance methods in hierarchical clustering	42
2.4	Linkage methods in hierarchical clustering	42
2.5	Cutting a dendrogram to assign clusters	43
2.6	K-means algorithm and scree plot	44
2.7	Microarray schematic	48
2.8	Network pruning and scale-free networks	52
2.9	WGCNA equations	53
2.10	PCA and LDA loadings	58
3.1	Schematic of data analysis workflow plan	65
3.2	GAinS: Visualisation of different linkage methods for hierarchical clustering	69
3.3	MOSAIC: Visualisation of different linkage methods for hierarchical clustering	70
3.4	GAinS: optimal cluster number determination	72
3.5	Clustered GAinS protein biomarker heatmap	74
3.6	GAinS: scaled, mean protein biomarker values in each cluster	75
3.7	MOSAIC k-means elbow and consensus cluster results	76
3.8	Clustered MOSAIC immune mediator heatmap	77
3.9	MOSAIC: scaled mean immune mediator values in each clusters.	79
3.10	HARP-2 optimal cluster determination	81
3.11	Principal component projection of HARP-2 protein biomarker values	81
3.12	HARP-2 biomarker heatmap and mean cluster levels	83
3.13	GAinS: Heatmap showing the correlations between protein biomarkers in septic patients	86

3.14	GAinS microarray normalisation boxplots	90
3.15	Pairwise scatter MDS plot for batch effects	91
3.16	GAinS: ARDS and non ARDS differential gene expression volcano plot . .	93
3.17	Soft power threshold and WGCNA module dendrograms for GAinS and MOSAIC microarray data	95
3.18	GAinS and MOSAIC eigengene module adjacency	101
3.19	GAinS: gene modules and patient traits correlation heatmap	103
3.20	MOSAIC: gene modules and patient traits correlation heatmap	106
4.1	GAinS: volcano plots for differentially expressed genes between patients with ARDS in each cluster	113
4.2	MOSAIC: volcano plots for differentially expressed genes between patients with severe respiratory failure in each cluster	118
4.3	MOSAIC: Lymphocyte and neutrophil counts in each cluster	120
4.4	GAinS: gene module and serum cytokine cluster correlation heatmap	121
4.5	MOSAIC: gene module and serum cytokine cluster correlation heatmap . .	122
4.6	Schematic of the GAIT mechanism. Taken from Mukhopadhyay et al. (2009)	123
4.7	MOSAIC: IFN- γ levels in each cluster	124
4.8	Figure 3b from Nie, Sun, Fun and Yu. Nature Cells and Disease 2019 depicting the role of EPRS in anti-viral immunity	126
4.9	GAinS MOSAIC: principal component plots for combined cytokine and eigenegenes values	128
4.10	GAinS MOSAIC: Three dimensional representation of combined cytokine and eigengene values.	129
4.11	GAinS: LDA projections and decision boundaries in ARDS	132
4.12	GAinS: Top ten ranked discriminators for ARDS samples between ‘purple’ and ‘green’ clusters	134
4.13	GAinS: Top ten ranked discriminators for ARDS samples between ‘yellow’ and ‘green’ clusters	136
4.14	GAinS: Top ten ranked discriminators for ARDS samples between ‘purple’ and ‘yellow’ clusters	138
4.15	GAinS: Metabase sub-networks from transcripts in ‘black’ and ‘dark orange’ modules	140
4.16	MOSAIC: LDA projections and decision boundaries in ARDS	146
4.17	MOSAIC: Top ten ranked discriminators for between ‘red’ and ‘grey’ clusters	148
4.18	MOSAIC: Top ten ranked discriminators between ‘blue’ and ‘red’ clusters .	150

4.19	Cartoon from London et al (2010) depicting the effect of Slit-2 ROBO-4 signalling on pulmonary endothelial leak	151
4.20	MOSAIC: Top ten ranked discriminators for ARDS samples between ‘blue’ and ‘grey’ clusters	154
4.21	HARP-2: LDA projection of biomarker data	156
4.22	Preliminary endotype schematic	162
5.1	GAinS: Boxplots comparing clinical variable measurements between patients from each endotype	167
5.2	GAinS: Boxplots comparing clinical variable measurements between patients with ARDS from each endotype	168
5.3	GAinS: Kaplan-Meier curve for 30 day survival of patients with each endotype	170
5.4	GAinS: Sankey diagram showing transitions between endotypes over three sampling times, alongside patient outcomes.	171
5.5	MOSAIC: Boxplots comparing clinical variable measurements for patients with $\text{rSOFA} \geq 2$ in each endotype	174
5.6	MOSAIC: Boxplots comparing clinical variable measurements for patients with $\text{rSOFA} \geq 3$ in each endotype	175
5.7	MOSAIC: Kaplan-Meier curves showing 30 day survival of patients with each endotype	179
5.8	MOSAIC: Sankey diagram of endotype transitions and hospital outcomes .	181
5.9	HARP-2: Boxplots comparing the clinical characteristics of patients in each endotype	186
5.10	HARP-2: 28 day survival curves for each endotype, stratified by treatment group	189
6.1	Final endotype model from each contributing study	196
D.1	GAinS: Boxplots of protein biomarkers concentrations in patients from each cluster	239
D.2	MOSAIC: Boxplots of protein biomarkers concentrations in patients from each cluster	243
D.3	HARP-2: Boxplots of protein biomarkers concentrations in patients from each cluster	244
E.1	PCA biplots for combined protein biomarker and eigengene data for each sample	245
F.1	GAinS: LDA rankings ‘purple’-‘green’	247
F.2	GAinS: LDA rankings ‘green’-‘yellow’	248

F.3	GAinS: LDA rankings ‘purple’-‘yellow’	249
G.1	MOSAIC: LDA rankings ‘blue’-‘grey’	251
G.2	MOSAIC: LDA rankings ‘blue’-‘red’	252
G.3	MOSAIC: LDA rankings ‘red’-‘grey’	253

List of tables

1.1	Randomised controlled trials in ARDS	6
2.1	Overview of the GAINs, MOSAIC and HARP-2 Studies	35
2.2	Respiratory component of the sequential organ failure (SOFA) score	40
3.1	Adjusted Rand index for missing data imputation strategies	67
3.2	GAINs: Mean z -scores of protein biomarkers in each cluster	73
3.3	MOSAIC: Mean z -scores of immune mediator in each cluster	78
3.4	HARP-2: Mean z -scores of protein biomarkers for each cluster	82
3.5	Stability of clusters in all three studies	84
3.6	GAINs: enrichment of WGCNA-derived gene modules	99
3.7	MOSAIC: enrichment of WGCNA-derived gene modules	100
4.1	HLH diagnostic criteria	144
4.2	MOSAIC: Transcripts from the antimicrobial humoral response gene module which discriminated the ‘red’ and ‘blue’ clusters	152
4.3	HARP-2: Ranked LDA coefficients for pairwise comparisons	155
5.1	GAINs: Clinical characteristics of patients from each of the three identified endotypes.	166
5.2	MOSAIC: Clinical characteristics of patients in each endotype	173
5.3	MOSAIC: Outcomes for patients in each endotype	178
5.4	HARP-2: Characteristics of patients in each endotype	185
5.5	HARP-2: Outcomes for patients in each endotype	188

Abbreviations

3' UTR	3 prime untranslated region
A(H1N1)pdm2009	Influenza A (H1N1) pandemic 2009
Ang-2	Angiopoietin-2
ANOVA	Analysis of variance
APACHE	Acute physiology and chronic health evaluation
ARDSnet	NIH-NHLBI ARDS research network
ARI	Adjusted Rand index
ARRB1	Beta arrestin-1
AT1	Angiotensin II type 1 receptor
AUROC	Area under the receiver operating characteristic curve
BALF	Bronchoalveolar lavage fluid
BPGM	Bisphosphoglycerate mutase
CC-16	Clara cell protein 16
CCL	C-C motif chemokine ligand
CDK5	Cyclin dependant kinase 5
cDNA	Complementary DNA
CMV	Cytomegalovirus
CSF1R	Receptor for colony stimulating factor 1
CXCL	C-X-C motif chemokine ligand
CXCR	Chemokine receptor
DAMPs	Damage associated molecular patterns
DBScan	Density-based spatial clustering of applications with noise
EBV	Epstein-Barr virus
ELISA	Enzyme-linked immunosorbent assay
ENG	Endoglin
<i>EPRS</i>	Gene which encodes glutamyl-prolyl-tRNA synthetase
eQTL	Expression quantitative trait loci

ER	Endoplasmic reticulum
FDR	False discovery rate
FiO ₂	Fraction of inspired oxygen
GAinS	Genomic advances in sepsis
GAIT	Gamma-interferon mediated inhibition of translation
GEO	Gene expression omnibus
GWAS	Genome-wide association study
HARP-2	HMG-Co-A reductase inhibitors in ARDS-2
HIV	Human immunodeficiency virus
HLH	Haemophagocytic lymphohistiocytosis
HRQoL	Health-related quality of life
I-TAC (CXCL11)	Interferon-inducible T cell alpha chemoattractant
ICAM-1	Intercellular Adhesion Molecule 1
IFN	Interferon
IFN- α 2a	Interferon alpha-2a
IFN- β	Interferon beta
IFN- γ	Interferon gamma
IFN- λ (IL-29)	Interferon lambda (interleukin 29)
IL-1 β	Interleukin 1-beta
IL-10	Interleukin 10
IL-12p70	Interleukin 12-p70
IL-13	Interleukin 13
IL-15	Interleukin 15
IL-17	Interleukin 17
IL-2	Interleukin 2
IL-23	Interleukin 23
IL-4	Interleukin 4
IL-5	Interleukin 5
IL-6	Interleukin-6
IL-8 (CXCL-8)	Interleukin 8
IP-10 (CXCL-10)	Interferon-gamma-inducible protein-10
<i>IRE-1</i>	Gene which encodes inositol-requiring enzyme 1 α
JAK	Janus kinase, a non-receptor tyrosine kinases
kME	Explained variance of the first principal component of gene module
LDA	Linear discriminant analysis
LPS	Lipopolysaccharide

MARS	Molecular Diagnosis and Risk Stratification of Sepsis
MCP-1 (CCL2)	Monocyte chemoattractant protein 1
MCP-4 (CCL13)	Monocyte chemoattractant protein 4
MDC (CCL22)	Macrophage-derived chemokine
MDS	Multi-dimensional scaling
ME	Module eigengene
MESSI	Molecular Epidemiology of Sepsis in the ICU
MHC	Major histocompatibility complex
MICE	Multiple imputation by chained equations
MIF	Macrophage inhibitory factor
MIG (CXCL9)	Monokine induced by gamma interferon
MIP-1 α (CCL3)	Macrophage inflammatory protein-1-alpha
MIP-1 β (CCL4)	Macrophage inflammatory protein-1-beta
MMP-8	Matrix-metalloproteinase-8
MOSAIC	Mechanisms of severe acute influenza consortium
MR	Mendelian randomization
mRNA	Messenger RNA
MSC	Multi-synthetase complex
NF κ B	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGAL	Neutrophil gelatinase-associated lipocalin
NHLBI	National heart, lung and blood institute (USA)
NIH	National institute for health (USA)
NK	Natural killer (cell)
OCLN	Occludin
OLS	Ordinary least squares
OPTICS	Ordering points to identify the clustering structure
PAI-1	Plasminogen activator inhibitor-1
PaO ₂	Partial pressure of arterial oxygen tension
PCA	Principal component analysis
PCR	Polymerase chain reaction
PCT	Procalcitonin
PECAM-1	Platelet endothelial cell adhesion molecule 1
PEEP	Positive end-expiratory pressure
PMVEC	Pulmonary-derived microvascular endothelial cell
PPI	Protein-protein interactions
PSGL1	P-selectin glycoprotein ligand-1 (CD162)

qPCR	(real-time) Quantitative polymerase chain reaction
RANTES (CCL5)	Regulated on activation, normal T cell expressed and secreted
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RRT	Renal replacement therapy
RSN	Robust spline normalisation
<i>S. pneumoniae</i>	<i>Streptococcus pneumoniae</i>
SAPS	Simplified acute physiology score
<i>SELPLG</i>	Gene which encodes the protein PSGL1
SOFA	Sequential organ failure assessment
SP-D	Surfactant protein-D
sRAGE	Soluble receptor for advanced glycation end products
SRS	Sepsis response signature
STAT	Signal transducer and activator of transcription
sTNFR-1	Soluble TNF receptor 1
SVD	Singular value decomposition
TARC (CCL17)	Thymus and activation regulated chemokine
TNF	Tumour necrosis factor
TOM	Topological overlap measure
TRALI	Transfusion-related associated lung injury
VEGF	Vascular endothelial growth factor
VFD	Ventilator free day
vWF	von Willebrand factor
WASP	Wiskott–Aldrich syndrome protein
WAVE	WASP-family verprolin-homologous protein
WCSS	Within cluster sum of squares error

CHAPTER 1

Introduction

1.1 Definition and features of acute respiratory distress syndrome

Acute Respiratory Distress Syndrome (ARDS) is a type of respiratory failure and patients present with tachypnoea, rapidly progressive respiratory failure and refractory hypoxaemia. The pathophysiology relates to pulmonary inflammation and neutrophil recruitment which increases endothelial permeability. A chain of events are then set in motion resulting in: pulmonary oedema, alveolar exudates, hyaline membrane formation, reduced lung compliance, increased physiological dead space and shunting. These processes cause worsening hypoxaemia and patients become less responsive to supportive measures including mechanical ventilation. Radiological investigations show diffuse bilateral opacification.¹ ARDS was first described by Ashbaugh et al. in 1967 and formally defined in criteria set out by the American European Consensus Conference (AECC) in 1994.^{2,3} This definition was revised in 2012 by the ARDS Definition Task Force, now referred to as the Berlin Definition, which stipulates that a diagnosis of ARDS requires the following to be met:⁴

- Acute respiratory failure (within 7 days of the onset of acute illness), or new and worsening respiratory symptoms
- Bilateral opacities on chest imaging not fully explained by lobar/lung collapse or pleural effusions
- Respiratory failure not fully explained by cardiac failure or fluid overload (which can be assessed by echocardiography)
- Hypoxia with a $\text{PaO}_2\text{-FiO}_2$ ratio <40 kPa

- Respiratory support with ≥ 5 cmH₂O positive end expiratory pressure (PEEP)

This definition categorised ARDS into three groups, determined by the degree of hypoxia measured by the PaO₂-FiO₂ ratio :

- Mild: PaO₂-FiO₂ ratio between 40 and 26.6 kPa, with ≥ 5 cmH₂O continuous positive airway pressure (CPAP) or PEEP
- Moderate: PaO₂-FiO₂ ratio between 26.6 and 13.3 kPa, with ≥ 5 cmH₂O PEEP
- Severe: PaO₂-FiO₂ ratio <13.3 kPa with ≥ 5 cmH₂O PEEP

The higher severity grades are associated with an increased risk of mortality, although their relative sensitivity for predicting mortality are somewhat lower than one might have expected given the severity of physiological derangement in these patients (AUC = 0.577).⁴

ARDS severity has previously also been categorised by using the lung injury score (LIS), sometimes referred to as the Murray score. This incorporates the degree of radiological changes (number of opacified lung quadrants on chest imaging) with three clinical features (static lung compliance, PaO₂-FiO₂ ratio and level of PEEP).⁵ This scoring system is recommended by the Extracorporeal Life Support Organisation (ELSO) as a tool to help select patients for extracorporeal membrane oxygenation (ECMO). ELSO recommends that patients with LIS ≥ 3 should be considered for ECMO. The score is a poor predictor of mortality during the first 72 hours following diagnosis of ARDS, although a score ≥ 2.5 between days 4 and 7 is associated with prolonged mechanical ventilation and a higher rate of in-hospital complications.⁴

1.2 Epidemiology and patient outcomes

ARDS is common, with approximately 190,000 cases and 74,000 deaths annually in the USA.⁶ Following the introduction of the Berlin definition, an international observation study (LUNG SAFE) was conducted in over 50 countries, involving 459 ICUs over four weeks in 2014. The investigators discovered that approximately 10.4% of ICU patients, and 23.4% of patients requiring mechanical ventilation satisfied ARDS criteria of at least mild severity, which often went unrecognised.⁷ Mortality for patients with severe ARDS during this period was 46.1% (95% CI 41.9%-51.4%), which was no improvement from that described 50 years ago, despite the advances in other areas of medicine. The ELSO registry report from 2016 showed that even when patients receive ECMO, the mortality for ARDS patients remains at 46%.⁸

ARDS can be categorised as primary (e.g. bacterial or viral pneumonia, direct lung trauma), or secondary in nature (e.g. non-pulmonary trauma, abdominal surgery, non-pulmonary sepsis). Primary and secondary causes are also referred to as direct or indirect by some authors. Neither the primary cause of ARDS nor the severity of hypoxaemia are independently associated with clinical outcome. Instead, the factors that are independently associated with increased mortality tend to be non-modifiable and include older age, active malignancy, haematological malignancy and non-pulmonary organ failure.⁶

For patients that survive the acute phase of ARDS, there lies ahead a recovery that may be complicated by long term critical illness. ARDS survivors often suffer from persistent physical, physiological and neurocognitive deficits that prevent recovery to their pre-morbid functional state. As many as 66% of survivors fail to recover their exercise capacity to pre-morbid levels, even 1-2 years after ICU discharge.^{9,10} Patients have impaired and delayed recovery of muscle function whilst the majority of patients show no long term pulmonary dysfunction. There is significant variation in the results of pulmonary function tests following recovery from ARDS; the consistently identified abnormality across multiple studies is mildly impaired diffusion factor, which is of uncertain significance. Cognitive impairment, post traumatic stress symptoms, anxiety and depression are recognised as common in survivors, and there is a growing need to address the ICU interventions and treatments that might be contributing to neuropsychological dysfunction.^{6,11}

Health-related quality of life scores (HRQoL) of ARDS survivors demonstrate consistent decrements in both physical and mental health domains. The results from patients with ARDS, over the first two years after ICU discharge, are similar to the results from the general ICU survivor population.¹² The health burden of ARDS on the families and caregivers (psychological, physical and financial) is not often recognised when reporting the epidemiology or outcomes for these patients as there is rarely longer-term (>5 year) follow up.^{9,13} Neuropsychological disorders contribute more to caregiver burden than physical impairment in this population.¹⁴ 31% of survivors who were previously employed never return to work, and 77% report lost earnings at five years.¹⁵ ARDS, therefore, poses a significant burden to society given the loss of productive economic output with respect to patients and those who care for them.

1.3 Pathophysiology

Acute inflammation affecting the alveolar-capillary membrane is the primary finding in ARDS. There is an increase in endothelial permeability, which is associated with acute

inflammatory mediators and neutrophil recruitment into the airspace resulting in pulmonary oedema. The combination of activated neutrophils and inflammatory exudate in the alveoli damages pneumocytes and inactivates surfactant causing distal airspace collapse with progressive loss of the available surface area for gas exchange. Inflammatory processes inhibit hypoxic pulmonary vasoconstriction that would otherwise regulate the pulmonary vascular tone to prevent shunting of deoxygenated blood into the systemic circulation. The resulting hypoxaemia is compounded by impaired pulmonary compliance, which means that higher airway pressures are required to maintain alveolar minute ventilation. Given that there is less healthy lung tissue available to take part in gas exchange (physiological dead space), alveolar ventilation falls causing accumulation of carbon dioxide and type II respiratory failure.

Efforts to address the abnormal arterial blood gas tensions of oxygen and carbon dioxide can be partially mitigated by adjustment of mechanical ventilation settings (PEEP, mean airway pressure, inspiratory time, minute ventilation) or prone positioning. Higher airway pressures and exposure of healthy lung tissue to higher energy forces are directly injurious to the lung tissue.¹⁶ These effects manifest as worsening inflammation at a cellular level, and as baro-(pressure), atelec-(repeated cycled closure and opening of lung units), volu-(alveolar stretch) trauma at a tissue level. Patients may develop pneumothoraces and bronchopleural fistulae which prolong the duration of mechanical ventilation. If higher tidal volumes are administered to increase minute ventilation and improve CO₂ clearance, there is exacerbation of pulmonary inflammation which is associated with worse survival.^{17,18}

Lung biopsies show histological changes described as ‘diffuse alveolar damage’, although only approximately half of patients with a diagnosis of ARDS have this finding at post mortem. Diffuse alveolar damage is characterised by hyaline membrane formation and pulmonary exudates that tend to be rich in neutrophils. ARDS exhibits significant heterogeneity; in a single patient’s lung tissue and between different patients with a diagnosis of ARDS, where only some of the above features may be apparent.¹⁹

Paradoxically, patients tend to die from multi-organ failure and not refractory hypoxaemia. Based on human experimental data, it has been suggested that in ARDS there is a failure of the lungs to maintain their immunomodulatory role; trapping activated neutrophils, other leucocytes and their mediators. These dysregulated immune components are able to break-through into the systemic circulation. Here they cause dysfunction of other organs, most commonly the kidneys, shortly followed by the cardiovascular system and liver. It has been shown that the lungs of patients with ARDS fail to deprime activated neutrophils, in contrast to the lungs of patients with sepsis and healthy controls.²⁰ Dysregulated immunological processes and multi-organ dysfunction are not unique to ARDS as they feature in patients with

common direct and indirect causes of ARDS (sepsis, trauma, major abdominal surgery, acute pancreatitis). Determining the dysfunctional immunological components that are pertinent to the pathophysiology of ARDS, but separate from other disease processes occurring in these patients, is therefore complex and difficult to demonstrate experimentally.

Patients who are successfully supported through ARDS may then progress to a prolonged resolution stage of their pulmonary disease.²¹ This stage is associated with slow resolution of pulmonary function, fibroblast proliferation, extracellular matrix and fibrin deposition and persistent shunting with associated hypoxaemia.¹ There is an increasing recognition that fibroproliferation occurs early in ARDS. High concentrations of N-terminal peptide for type III procollagen (N-PCP-III) in both the sera and bronchoalveolar lavage fluid (BALF) at 24 hours in ARDS patients are associated with a worse outcome.²² Prolonged periods of mechanical ventilation increase the risk of developing other complications of intensive care or hospital treatment (e.g. secondary infections, delirium, venous thromboembolism, pressure sores, myopathy and neurocognitive dysfunction).²³

1.4 Treatment of ARDS

There have been many multi-centre trials aimed at improving the outcomes for patients with ARDS (Table 1.1). Of all the interventions studied there have been only two supportive measures and two pharmacological therapies that improved outcomes for patients:

- Low tidal volume ventilation (6 mL/kg predicted body weight)¹⁷
- Cis-atracurium, a non-depolarising neuromuscular blocking drug that prevents contraction of skeletal muscles²⁴
- Prone positioning²⁵
- Dexamethasone, a steroid drug²⁶

The DEXA-ARDS trial showed an improvement in outcomes for patients randomised to dexamethasone but the trial was stopped early at 88% planned recruitment (288/314 patients) by the data safety monitoring board. This was because of the low enrolment rate of this study. The authors reported significant benefits for patients receiving dexamethasone with respect to improvement in ventilator free days (4.8 day reduction [95% CI 2.57 to 7.03]; $p < 0.0001$) and 60-day mortality (absolute risk reduction -15.3% [95% CI -25.9 to -4.9]; $p = 0.0047$). However these are partial results due to the early cessation of the trial and must be interpreted with caution.²⁶

Study	Publication	n	Principal Intervention	Primary Outcome
ARDSnet:ARMA ¹⁷	2000	861	Low tidal volume ventilation	Improved hospital mortality
ARDSnet:KARMA ²⁷	2000	234	Ketoconazole	No difference in 28 day mortality
ARDSnet:LARMA ²⁸	2002	235	Lisofylline	No difference in 28 day mortality
ARDSnet:ALVEOLI ²⁹	2004	549	High PEEP strategy	No difference in hospital mortality
<i>Taylor et al (JAMA)</i> ³⁰	2004	385	Nitric Oxide	No difference in VFD at 28 days
ARDSnet:LaSRS ³¹	2006	180	Methylprednisolone	No difference in 60 day mortality
ARDSnet:FACTT ³²	2006	1000	Liberal or conservative fluid strategy	No difference in 60 day mortality
ACURASYS ²⁴	2010	340	Cis-atracurium	Improved 90 day mortality
ARDSnet:Omega ³³	2011	272	Omega-3 fatty acid	Stopped early for futility
ARDSnet:ALTA ³⁴	2011	282	Albuterol (inhaled)	No difference in VFD at day 28
BALTI-2 ³⁵	2012	162	Salbutamol (IV)	Increased mortality, stopped early
ARDSnet:EDEN ³⁶	2012	1000	Trophic vs Full enteral nutrition	No difference in VFD at day 28
HARP-2 ³⁷	2014	540	Simvastatin	No difference in VFD at day 28
OSCAR ³⁸	2013	795	HFOV	No difference in 30 day mortality
OSCILLATE ³⁹	2013	548	HFOV	Increased mortality, stopped early
PROSEVA ²⁵	2013	466	Prone positioning	Improved 28 day mortality
ARDSnet:SAILS ⁴⁰	2014	745	Rosuvastatin	Stopped early for futility
LIPS-A ⁴¹	2016	390	Aspirin	No difference in the incidence of ARDS
KARE ⁴²	2017	268	Recombinant keratinocyte growth factor	No difference in VFD at day 28, higher mortality in intervention group
EOLIA ⁴³	2018	247	ECMO	No difference in 60 day mortality
ROSE ⁴⁴	2019	1006	Cis-atracurium	Stopped early for futility
DEXA-ARDS ²⁶	2020	277	Dexamethasone	Improved VFD and 60 day mortality

Table 1.1 Randomised controlled trials in ARDS VFD - ventilator free days. PEEP - positive end expiratory pressure. HFOV - high frequency oscillatory ventilation. ECMO - extracorporeal membrane oxygenation

The lack of precise biological mechanisms to target therapies has been a persistent theme in ARDS. Patient heterogeneity and limited characterisation of patient features, using the available diagnostic criteria, only compound the noise in the data when studying this condition. In heterogeneous syndromes there will be members of the patient population who benefit from a particular intervention, others may come to harm whilst others experience no change to their outcome. The consequences of being unable to determine which patients might benefit from a particular set of interventions is that these treatments are discarded resulting in the persistent high mortality for those with severe disease.⁷

ARDS is not unique in having this repeating pattern of unsuccessful interventions. Randomised controlled trials (RCTs) in other critical illnesses syndromes (acute kidney injury, sepsis, cardiogenic shock) have suffered a similar fate.^{45–48} Each of the promising interventions and therapies in these organ failure syndromes, based on robust physiological reasoning or disease models, were expected to deliver improved patient outcomes. The value of these ‘negative studies’ is that they may prevent unnecessary, harmful treatments for our patients. Examples of this include doxycycline- α which was found, after marketing, to be associated with increased mortality.⁴⁹ Use of hydroxyethyl starch-based fluids are associated with acute kidney injury in patients with sepsis.⁵⁰ However, there may have been sub-populations of ICU patients that would have benefited from the treatments in each of these failed RCTs. These trials may have befallen a Type II statistical error by failing to apply their interventions to the correct patients.

It is increasingly apparent that critical illness is a collection of poorly characterised, heterogeneous clinical syndromes rather than distinct diseases. Our current, routine biochemical tests and physiological measurements are unable to differentiate the nuances between different subtypes. Avoiding repetition of the failures over the past 50 years will require patient stratification to determine which groups of patients may or may not benefit from a novel therapeutic intervention. This problem is not unique to critical care. Still, given there are no definitive, diagnostic biomarkers for many critical illnesses (ARDS, sepsis) the diagnostic uncertainty in these patients compounds the potential errors.

In light of the many discarded treatments, there is an urgent need for a tailored approach. The goal is to characterise patients accurately into disease subtypes by incorporating genomic and our understanding of biological processes with the physiological response to acute illness. This individualisation of care is often referred to as precision or stratified medicine. The expectation is the ability to seamlessly integrate multi-modal information to individualise therapy and change the disease trajectory.

1.5 Stratification approaches in ARDS

Methods that have been used to stratify patients with ARDS can broadly be classified into:

- Biomarkers
- Genomics
- Physiology

1.5.1 Biomarkers in ARDS

The FDA define a biomarker as:

“A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers. A biomarker is not an assessment of how a patient feels, functions, or survives.”

*BEST (Biomarkers, EndpointS, and other Tools) Resource.
FDA-NIH Biomarker Working Group 2016.*

Biomarkers are classified according to the role they are being used for. In ARDS, investigators have focused on diagnostic biomarkers and prognostic biomarkers. Accurate biomarkers are necessary for ARDS as the Berlin definition (and its predecessors), which are based on clinical measurements, are poor predictors of patient outcomes.⁴ Diagnostic biomarkers determine the presence or absence of a disease or disease subtype (e.g. sweat chloride in cystic fibrosis). Prognostic biomarkers indicate the likelihood of a future clinical event in an identified population (e.g. prostate-specific antigen and likelihood of prostate cancer progression). Predictive biomarkers are used to identify individuals that might respond differently to a given treatment or environmental exposure (e.g. thiopurine methyltransferase genotype or activity and risk of toxicity from azathioprine). An identified biomarker may not attribute a mechanism to the disease in question and so positive correlations or associations must be interpreted with appropriate caution.

In ARDS, diagnostic and prognostic biomarkers have been sought by measuring cytokines and chemokines in serum and sampled lung washings (BALF). The search for the ARDS

equivalent of a high sensitivity troponin following myocardial injury remains elusive. Due to the severe hypoxaemia in patients with ARDS, sampling the lungs is not feasible in large observational studies. The biomarkers measured in such studies are from blood samples and can be classified as pulmonary-derived, vascular-derived or cytokines.

Pulmonary

Pulmonary biomarkers are derived from pulmonary epithelial tissue. These proteins may be released into the circulation following injury to alveolar type I or alveolar type II cells.

- soluble receptor for end glycosylation products (sRAGE) is strongly expressed in pulmonary epithelium, especially alveolar type I cells. sRAGE plasma levels in patients with severe ARDS have been shown to correlate with mortality. High levels of detectable plasma sRAGE are not specific to ARDS as other pulmonary and non-pulmonary diseases are associated with raised levels.⁵¹
- Surfactant protein D (SP-D) is one of four surfactant proteins produced by alveolar type II cells. Raised levels correlate with ARDS mortality and levels tend to be higher in direct ARDS.⁵¹
- Krebs von den Lungen-6 (KL-6, now officially named Mucin-1) is associated with mortality in ARDS. It is a large glycoprotein expressed on type II alveolar cells and is associated with lung inflammation. Raised levels are associated with mortality in ARDS patients.⁵¹

Vascular

Vascular biomarkers have a role in endothelial function or coagulation. These include angiopoietin-2 (Ang-2), von Willebrand factor (vWF) and plasminogen activator inhibitor-1 (PAI-1).

- Angiopoietin-2 is one of a family of growth factors stored and secreted by endothelial cells. They affect vascular permeability and remodelling of vascular tissue. Raised circulating Ang-2 also associated with sepsis, diabetes and both solid and haematological malignancies.⁵² Raised Ang-2 in both ARDS and at risk patients are predictive of mortality and correlate with ARDS development in patients with severe traumatic injuries.^{53,54}
- Von Willebrand factor is an important glycoprotein involved in coagulation. It binds coagulation factors, platelets and endothelial cells to help achieve haemostasis fol-

lowing vascular injury. Raised vWF levels are associated with sepsis and ARDS. Ware et al. (2004) found there was no difference in concentrations between patients with non-septic and sepsis-associated ARDS. However, higher concentrations were associated with mortality in these patients with ARDS. The authors also found that vWF concentrations were significantly lower in patients with trauma associated ARDS compared with ARDS related to other causes.⁵⁵

- Plasminogen activator inhibitor-1 (PAI-1) is protein released by endothelial cells that is responsible for the inhibition of tissue plasminogen activator (tPA) and urokinase. Both of these two enzymes are responsible for activation of plasmin which catalyses the degradation of fibrin, the primary protein constituent of blood clots. PAI-1 therefore promotes fibrin integrity. High levels of PAI-1 are found in both the alveoli and plasma of patients with ARDS, and are associated with higher mortality.⁵⁶ This protein may play a role in alveolar fibrin deposition and inappropriate activation of coagulation vascular microthrombosis in pulmonary capillaries, both of which of which are histopathological features in patients with ARDS.¹⁹

Cytokines

Cytokines are proteins that play a role in immunological signalling. IL-2, IL-4, IL-6, IL-8, IL-1 β , TNF- α have all been associated with ARDS and mortality. Their overlap with sepsis and other inflammatory states (trauma, burns) makes them less useful as single predictors in ARDS. A recent meta-analysis indicated that IL-8 was the most strongly associated with the diagnosis of ARDS (of the above cytokines), whilst IL-2 and IL-4 were strongly associated with mortality.⁵⁷

Do combinations work better?

Ware et al. (2010) developed a predictive model which used a combination of physiological features (APACHE III, age, number of organ failures, alveolar-arterial oxygen difference, age) with eight biomarkers (SP-D, vWF IL-6, IL-8 TNFR-1, PAI-1, ICAM-1 and protein C) in 528 patients to predict mortality in ARDS.⁵⁸ They developed models in both sepsis and trauma-associated ARDS, using patients without ARDS as controls. The strongest predictors that were common to both the trauma and sepsis-associated ARDS were the combinations of APACHE-III, IL-8 and SP-D (AUC = 0.834). This approach was validated in a further paper (Zhao et al. 2017) using the same predictors in 1538 patients; the model performed well (AUC = 0.74), but not quite as well as their original 2010 paper (AUC = 0.85).⁵⁹ The same group has also used similar a strategy for diagnosis of ARDS where, using 100 patients, developed a five biomarker panel (SP-D, sRAGE, IL-8, CC16 and IL-6) for diagnosing ARDS in sepsis (AUC = 0.75).⁶⁰

Calfee et al. (2015) showed that measurements of single proteins performed adequately as predictive and diagnostic biomarkers. By comparing the relative concentrations of different cytokines in patients with direct (pneumonia, aspiration, thoracic trauma) and indirect (sepsis, trauma) ARDS, they found raised SP-D and sRAGE to be significantly higher in patients with direct ARDS, and Ang-2 to be significantly higher in patients with indirect ARDS. These results were validated from samples collected and analysed *post hoc* from a multi-centre study of patients (n = 853). Their findings were consistent with the association of direct ARDS with pulmonary epithelial injury, and non-direct ARDS with endothelial inflammation. Unlike other attempts to find predictive signals in ARDS, this approach described differential biological processes, albeit crudely, compared with a purely statistically-driven, model optimisation approach.⁶¹

1.5.2 Genomic studies in ARDS

There have been six human transcriptomic studies (using microarrays), two genome-wide association studies (GWAS) and three focused genome association studies in ARDS.

Transcriptomics

The largest transcriptomic study to date was by Sweeney et al. (2018) who investigated whole blood gene expression in 148 patients ARDS and 268 controls. They found a set of 30 differentially regulated genes, which enriched for expression in metamyelocytes and granulocytes. The authors attributed this signal to inflammatory and non-pulmonary processes. A seven-gene subset of the thirty performed poorly and had low generalizability for diagnosis of ARDS. The authors proposed no mechanisms based on these results.⁶²

Kangelaris et al. (2015) investigated differential gene expression between patients with ARDS and sepsis controls ($n = 57$), creating two models, one of which was adjusted for age, sex, batchⁱ, type of ARDS (direct / indirect) and neutrophil counts. Fifteen genes were differentially expressed, of which four were consistently upregulated. qPCR was performed to confirm the higher expression levels of these genes. One of these genes was CD24, the granulocyte receptor for platelet P-selectin, which is involved in platelet-neutrophil interactions and was later identified in a genome-wide association study (GWAS) by Bime et al. (see below). The other three genes were: lipocalin-2 (also known as NGAL), bactericidal permeability-increasing protein (BPI) and neutrophil collagenase (MMP-8), all of which are associated with neutrophils.⁶³

Howrylak et al. (2009) examined whole blood gene expression in 13 patients with ARDS and 21 sepsis controls. They found eight differentially expressed genes, the strongest of which was the ferritin heavy chain. The role of raised ferritin was presumed to be a sign of oxidative stress in ARDS patients, however, their results did not suggest any further mechanistic insights and was limited by a small sample size.⁶⁴

Chen et al. (2013) used the same data collected by Howrylak et al. (2009) submitted to the gene expression omnibus (GEO) repository. Using updated informatics methods they identified twenty differentially expressed genes (12 upregulated, 8 downregulated). Following enrichment, the authors focused on occludin (OCLN) and HLA-DQB1. OCLN is a membrane protein involved in tight junction assembly which may be influenced by TNF- α and IL-18 signalling. HLA-DQB1 is a major histocompatibility complex (MHC) class II protein.

ⁱ‘Batch’ refers to non-biological factors, usually environmental or technical conditions that change experiment results. High throughput genomics experiments are particularly susceptible to batch effects due to variability in the chemistry of the reaction steps and sensitivity to detect small changes in gene expression.

MHC proteins present antigens to immune cells and so play an integral role in the immune system.⁶⁵

Dolinay et al. (2012) performed whole blood gene expression on 88 patients with sepsis and ARDS. The study focused on validating the role of inflammasomes and IL-18 in ARDS using a mouse model. Results from the gene expression data confirmed the high expression of inflammasome related genes (caspase-1 and ASC) in ARDS compared with sepsis patients. Kangalaris et al. attempted to replicate these findings but found the relative expression of IL-18 was lower in the patients recruited to their study.^{63,66}

Juss et al. (2016) compared the transcript profiles blood neutrophils in ARDS patients with healthy volunteers (n = 12) and found 1319 differentially expressed genes. This list was refined to 216 differentially expressed genes when the healthy volunteer neutrophils were GM-CSF treated. There was an interesting overlap between the upregulated, differentially expressed genes from these neutrophils and the results from a study comparing leucocyte gene expression in burns patients with healthy volunteers.⁶⁷

Genome-wide association studies

Two GWAS have been published in ARDS. Christie et al. (2012) identified 159 enriched single nucleotide polymorphism (SNPs) in 812 patients following major trauma (600 discovery, 212 validation).⁶⁸ One locus, *PPFIA1*, was significant following expression quantitative trait loci (eQTL) analysis of a B-lymphoblastoid cell line. This result was of nominal statistical significance and no polymorphism had genome wide significance. *PPFIA1* encodes liprin- α which is involved in cell adhesion and cell-matrix interactions.

The second GWAS, conducted by Bime et al. (2018), recruited 232 African-American patients with ARDS and the authors identified an intragenic SNP in *SELPLG* to be associated with increased susceptibility of developing ARDS. This gene encodes P-selectin glycoprotein ligand-1 (PSGL1). The role of this gene in ARDS was demonstrated using murine models of ventilator induced and lipopolysaccharide (LPS)-induced lung injury. The authors found increased expression in these mice, and attenuation of response after treatment with neutralising antibodies.⁶⁹ This protein had previously been identified by Kangalaris et al. (2015) which offered some validity to this finding.⁶³

Although both GWAS highlighted some potential insights into ARDS their results were of borderline statistical significance; the *p* value for rs471931, a *cis*-acting SNP influencing the differential expression *PPFIA1*, was of low significance. In order to identify the SNP rs2228315 in *SELPLG*, the *p* value threshold had to be lowered from the normal Bonferroni

corrected threshold of $p < 3.5 \times 10^{-8}$ to $p < 10^{-3}$. Both of these studies can therefore only be considered hypothesis generating.

Focused genomic association studies

Three focused genome association studies have identified three loci associated with ARDS. Hernandez-Pachecho et al. (2018) integrated transcriptomic analysis of a murine ARDS model with Christie et al.'s (2012) GWAS study to identify prioritized genes that might be important in ARDS.^{68,70} Four candidate genes were identified. One SNP (rs9513106 in *FLT1*) was associated with a lower risk ARDS in a Spanish intensive care genotyped cohort of 1,124 individuals. A validation cohort from a GWAS of 2,355 individuals from the USA confirmed these results with statistical estimates of effect size that were similar to that found in the discovery cohort. A meta-analysis of both studies found the *C* allele at this locus in *FLT1* was associated with an odds ratio equal to 0.77 (95% CI 0.65-0.92; $p=0.003$) for sepsis-induced ARDS.

FLT1 encodes a tyrosine kinase receptor with an extracellular ligand-binding region containing seven immunoglobulin-like domains. FLT1 belongs to the vascular endothelial growth factor (VEGF) receptor family. These domains can bind VEGF-A, VEGF-B and other growth factors that increase endothelial permeability, causing tissue oedema. High levels of VEGF have previously been identified in the bronchoalveolar lavage fluid of patients with ARDS.⁷¹ The function of this protein and the prospect of polymorphisms in this gene influencing the risk of ARDS were therefore plausible.

The second of these three focused genome association studies was published by Reilly et al. (2018).⁷² The investigators genotyped 703 patients with sepsis as part of the Molecular Epidemiology of Sepsis in the ICU (MESSI) study. The genomic analysis was limited to a 70 kilobase region around the *ANPT2* gene which encodes the protein angiopoietin-2 (Ang-2), a marker of endothelial activation previously associated with poor outcomes in ARDS (Section 1.5.1). SNPs that were associated with raised levels of Ang-2 were then tested for associated with ARDS. The authors undertook two approaches to verify their findings: Mendelian randomization and mediation analysis.

Mendelian randomization (MR) uses genetic mutations as fixed instruments to assess the causal effect of observed associations (plasma Ang-2) with outcomes (ARDS risk). The model incorporates adjustments for confounding variables that might influence both the observed variable and the outcome. For example, in this study the authors used APACHE-III scores, pulmonary source of infection and genetic ancestry as confounders. This methodology allows for assessment of the effects of genetic mutations on outcomes, via an instrumental

variable, in observational studies. Mediation analysis uses intermediate variables (plasma Ang-2) to quantify the causal effect of explanatory variables (SNP alleles) on the outcome (ARDS).

The authors demonstrated the association of five unlinkedⁱⁱ SNPs in patients of European ancestry with plasma Ang-2 concentrations. Of these, two (rs2442608 and rs2442630) were associated with ARDS, with rs2442630 having the largest effect size (OR = 1.38, 95% CI 1.01-1.87; $p = 0.04$). The MR analysis calculated that the genetically predicted Ang-2 concentration, as determined by the collective effects of the five identified SNPs, significantly increased the risk of ARDS (adjusted OR = 2.25, 96% CI 1.06-4.78; $p = 0.035$).

These findings were consistent with the earlier biomarker studies of Ang-2 and their results suggested that this protein and its effects might be an important therapeutic target for drug development. The authors were unable to demonstrate a causal effect of these SNPs in patients of African ancestry, even though raised Ang-2 concentrations in these patients was associated with ARDS. Ang-2 was associated with ARDS in patients with pulmonary or non-pulmonary sepsis. This was not consistent with the findings of Calfee et al (2015)'s biomarker study which found higher levels of Ang-2 in patients with indirect (non-pulmonary) causes of ARDS.⁶¹ This inconsistency probably reflects the inter-observer variation in diagnosing patients with ARDS.

An earlier study by Gong et al. (2007) used a nested-case control model to investigate the role of SNPs in a single gene (*MBL2*).⁷³ *MBL2* encodes the protein mannose binding lectin-2 which had been previously associated with increased susceptibility to severe to bacterial (meningococcal, pneumococcal) and viral (hepatitis B, severe acute respiratory syndrome caused by coronavirus) infections. The authors investigated four SNPs associated with this gene, three in exon 1 and one in an adjacent promoter region. 752 ICU patients, of which 237 developed ARDS, were genotyped at these four loci. Patients with homozygous alternative alleles of SNP rs1800450 (codon 54 of *MBL-2*) were associated with significantly worse multi-organ dysfunction, higher APACHE-III scores, increased risk of ARDS and higher mortality, compared with individuals who were homozygous or heterozygous for the reference allele. The frequency of the minor allele of this haplotype in the studied population was 14%. There was no secondary validation of these results. The authors attributed the worse outcomes of patients with the *BB* allele at *MBL-2 codon 54* to increased susceptibility to infection.

ⁱⁱnot in linkage disequilibrium

The results from candidate gene studies have to be interpreted with caution as they have been shown to be at risk of producing false positive results which can not be replicated despite compelling, statistically significant associations.⁷⁴ There are a number of reasons why false positive associations may arise from these studies:

1. The candidate genes may have been selected as part of a larger search for association between SNPs and disease features, but the results were only reported with respect to a candidate gene or genomic region. Casting a wide net and then only reporting the significant associations from a small number of genes would reduce the multiple correction penalty. Calculated *p* values would therefore appear smaller and more significant, when they may have arisen by chance alone.
2. Selection of the genomic region for analysis is subject to the investigators discretion. Inclusion of neighbouring promoter and silencer regions, cis-acting response elements, the local topologically active domain which might contain additional epigenetic targets that might associate with the candidate gene and influence its expression may or may not be included as part of these studies.
3. The biological plausibility of a candidate gene is an inadequate prior and does not provide adequate assurance against a false positive finding.⁷⁵
4. The definition of replication is subject to interpretation as there is a relative hierarchy in the quality of replication studies. A relaxed approach would permit any SNP with any direction of association with phenotype using any statistical test. A more stringent approach might require the exact same SNP, with same direction of association and the same statistical test, would help to eliminate false positive associations in the validation set.⁷⁵
5. A negative finding from a candidate gene study does not exclude it from a genuine association. Some associations have been found to be significant following meta-analysis. However, negative studies are often subject to a negative publication and time-lag biases which introduces a delay for these studies to be incorporated into a meta-analysis.⁷⁴

1.5.3 Clinical variable-based subtypes of ARDS

Attempts to score, stratify or predict outcomes in ARDS patients based on clinical variables have not been shown outperform existing scoring systems (APACHE-III, SAPS, SOFA). It is widely acknowledged that the Berlin definition is a relatively poor predictor of outcomes.^{4,76,77}

Villar et al. (2015) described four phenotypes based on a combination of PaO₂-FiO₂ ratio <150 mmHg (20 kPa) and PEEP >10 cmH₂O. This differs from the Berlin definition which stratifies patients with PaO₂-FiO₂ ratio 300 mmHg (40 kPa), 200 mmHg (26.7 kPa) and 100 mmHg (13.3 kPa) whilst receiving PEEP ≥ 5 cmH₂O. This method was more predictive of mortality if applied at 24 hours after ICU admission. Groups with lower PaO₂-FiO₂ ratio and higher PEEP requirements had a significantly higher mortality, but they also had higher APACHE scores and incidence of multi-organ failure. The authors acknowledged this in their discussion but did not incorporate these additional covariates into their analysis or make adjustments for them. Bos et al. (2016) used this scoring system in Dutch patients with ARDS and found the phenotype groups to align with 30-day mortality.^{78,79}

From a patient perspective, an important study by Wang et al. (2014) found that whilst acute physiological derangement predicted short term outcomes (ICU / 30 day mortality), these factors failed to predict, the more patient-focused, 1 year mortality outcome. In their multivariate model, the strongest predictors of 1 year survival were age, comorbidities and discharge destination. They found patients admitted to ICU had a 24% hospital mortality, reflecting the improvements in supportive care, but 41% 1 year mortality.⁸⁰

Each of the methods described above (biomarkers, genomics and physiology) are limited by recurring themes of low predictive validity, poor choice of control populations and comparator groups. Low predictive validity arises where the study findings can not be externally validated with other data, prospectively or, for the most part, yield no new mechanistic insights into the disease. A recurring issue with these studies is that patients with sepsis were the control cohort. There has been little acknowledgement of the inherent heterogeneity within sepsis itself. Treating the results from patients with sepsis as a statistically static control seems inherently flawed. These patients cannot be considered as equivalent to controls for an *in vitro* experiment or animal model. Adjustment for clinical covariates or stratification using clinical variables will not mitigate this heterogeneity if they do not reflect the differences in the underlying biological processes. Patients with different diseases or syndromes (cardiogenic shock, pneumonia, acute liver failure, major haemorrhage),

characterised by different biological processes, can arrive at the same physiological endpoints e.g. $\text{PaO}_2\text{-FiO}_2$ ratio < 100 mmHg, vasoplegic shock, high SOFA score.

Bos et al. (2019) have recently been able to demonstrate ARDS endotypes within a heterogeneous sepsis population. These patients had been recruited to an observational study of septic called the Molecular Diagnosis and Risk Stratification of Sepsis (MARS).⁸¹ In this study patients with sepsis from the MARS cohort, with and without ARDS were characterised into hyper-reactive and hypo-reactive subtypes based on their cytokine profiles. This study was an extension of the authors' previous work on endotyping ARDS using hierarchical clustering.⁸²

Differential gene expression, from whole blood transcriptome microarray analysis, between the patients in these two subtypes, revealed a number of processes that might be at work to differentiate non-inflamed and reactive ARDS. However, this study failed to acknowledge the degree of overlap between the sepsis patients and ARDS patients which was apparent in Figure 4 of their paper (Figure 1.1).

In this figure, the differential gene expression results from healthy controls, patients with sepsis, reactive and non-inflamed ARDS are shown, projected onto the first two principal component axes. K-means clustering was used to fit a medioid to each group of patients. These medioid locations are shown as coloured squares. Each medioid square was used to summarise the differentially expressed genes in non-inflamed ARDS, reactive ARDS and sepsis (as a single entity) from healthy controls.

Across both of the presented principal components in Figure 1.1, the gene expression profiles of septic patients (grey coloured points) spans from the healthy control population (black coloured points) to the reactive ARDS patients (red coloured points), and overlaps the uninflamed ARDS patients (green coloured points). The first principal component was used by the authors to determine the contributing pathways differentiating these groups of patients from each other. This figure captures the biological heterogeneity, at a gene expression level, in patients with sepsis or ARDS and demonstrates why direct comparisons between these two critical illness syndromes are difficult to interpret. Collapsing all of the heterogeneity amongst the grey points in this figure into the medioid represented by the grey square, which approximated closely to the medioid represented by the uninflamed ARDS patients (green square) might suggest that the uninflamed ARDS patients are similar in their gene expression profiles to patients with sepsis. This interpretation would be incorrect given the high variance, within the first principal component, of gene expression profiles in septic patients.

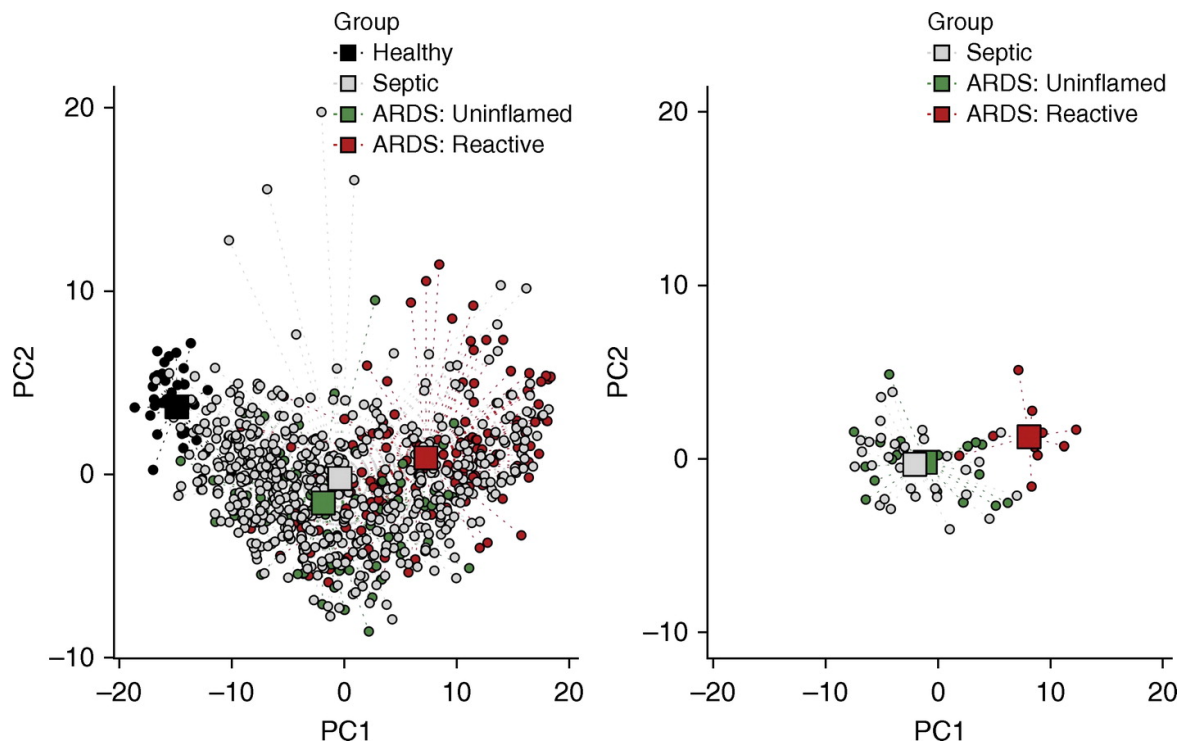


Fig. 1.1 Figure from Bos et al. (2019) showing gene expression PCA of patients with sepsis and ARDS subtypes. Left and right panels show the profiles from discovery and validation cohorts respectively. The x-axis shows the first principal component explaining around 56% of variance of the most differentially expressed genes in the derivation cohort. The y-axis shows the second principal component explaining around 8% of variance. Dots represent individual patients with colours identifying groups. The larger rectangle shows the location of the mediod for each group. The two ARDS subtypes overlies the septic patients and their separate mediods are plotted.

The apparent variation within the septic patients has not been accounted for, with only a single mediod for the entire spectrum of disease plotted. Taken from Figure 4 of Bos et al. (2019).⁸¹

Reprinted with permission of the American Thoracic Society.

Copyright © 2020 American Thoracic Society. All rights reserved.

Lieuwe D. J. Bos, Brendon P. Scicluna, David S. Y. Ong et al.

Understanding Heterogeneity in Biologic Phenotypes of Acute Respiratory Distress Syndrome by Leukocyte Expression Profiles.

AJRCCM 200(1) 2019 42-50.

The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society.

1.6 Successful endotyping approaches in other diseases

Diseases that might have previously been well characterised by their symptoms, natural history, clinical signs and investigations are now being recognised as collections of heterogeneous variants with different underlying pathophysiology. Examples include asthma, chronic obstructive pulmonary disease (COPD) and breast cancer. The opportunity to better describe these syndromes has emerged due to complementary improvements in scientific methods (genomics, high sensitivity assays) and computational methods (high dimensional data analysis, neural networks) all of which are increasingly accessible and have led to the new field of systems biology.

It is now possible to combine data from high-throughput biological methods with symptomatology and treatment information to generate models that identify novel patient sub-clusters. The characteristics that these specific subgroups share are better enriched for mechanisms compared with analysis of observable clinical features and characteristics in isolation. This might explain why genome-wide association and genetic linkage analysis of twins and family studies have not revealed new insights into clinically heterogeneous respiratory diseases like asthma or COPD outside of rare subsets (e.g. α -1 antitrypsin deficiency). Phenotypes that are characterised by specific pathobiological mechanisms are referred to as ‘endotypes’. As an illustrative example, sickle cell disease could be considered an endotype of anaemia.

1.6.1 Asthma

Asthma has long been recognised as a highly heterogeneous syndrome with subsets of patients that have remained treatment resistant. Following a number of major epidemiological studies and patient symptom-driven surveys to which latent class and other clustering methods have been used, new insights have emerged, pathological processes characterised and therapies developed.

The efforts of investigators to discover endotypes in asthma has involved use of different clustering methods. Loza et al. (2016) used partitioning around medoids (PAM) clustering on two asthma studies (ADEPT and U-BIOPRED) with training cohort sizes of 156 and 82, respectively. Validation steps were both longitudinal (ADEPT) and in a larger U-BIOPRED cohort ($n = 397$).⁸³ Four described phenotypes, which incorporated the biological data and patients symptoms were longitudinally stable, and biologically distinct.

Siroux et al. (2011) used latent class analysis (LCA) clustering on data obtained from two other asthma cohorts EGEA2 ($n = 641$) and ECRHSII ($n = 1895$). The authors identified

four phenotypes that shared similar features across both cohorts.⁸⁴ The authors found the LCA methods tended to relegate the importance of allergic symptoms. These results could be interpreted as clustering methods removing traditional clinical biases from the accepted disease model. Each identified phenotype had distinct health-related quality of life (HRQoL) scores and biological features. The identified clusters also overlapped with other studies.^{85,86}

There have been multiple iterations of latent class methods on patients symptoms, physiology and blood tests to determine asthma endotypes. These have further been extended into transcriptomic and biomarker-based studies to reveal the underlying processes e.g. T_H2 and non-T_H2 mediated asthma.^{87,88}

Hinks et al. (2016) used topological data analysis and Bayesian belief networks to identify six asthma phenotypes in 194 patients. These phenotypes were determined using data from assessment of symptoms, treatment response, physiological and protein biomarkers from sputum and blood. Validation used a separate 106 patient cohort where four of these six clusters were replicated, the other two clusters were small and their absence attributed to overfitting in the derivation cohort. The degree of inflammation was not found to correlate with clinical features which reiterated the theme of clinically-based phenotypes failing to stratify patients correctly.⁸⁹

1.6.2 Breast cancer

Molecular features (hormone receptors, HER2 status) have long been associated with breast cancer outcomes. Dawood et al. (2011) developed a composite multi-variable model of immunohistochemical features which they validated in a large cohort of 1,957 patients. They found five distinct tumour phenotypes of which 'luminal A' had a worse prognosis than the other four.⁹⁰ The advantage of this approach was that it could be applied to preserved histological samples, without the need for gene expression profiling using microarrays. The same group also found that ductal carcinoma in situ (DCIS) displayed the same five phenotypes found in invasive breast cancer but at different relative frequencies in a sample of 2897 patients.⁹¹

The success of these approaches in oncology has led to the development of platform trial design. Platform trials evaluate multiple treatments in a heterogeneous population and assume that treatment effects might also be heterogeneous. Treatment groups can change over time as data from the study is evaluated, and may even be dropped if there is evidence of harm or futility. The advantages of this approach are that they enrich the treatment groups for patients

that might be more likely to benefit from a particular treatment or intervention (response-adaptive randomisation). Lower numbers of patients are necessary for each treatment group and unsuccessful interventions can be discarded earlier. These approaches might therefore be more economical for sponsors whilst reducing the number of potentially exposed patients, unnecessarily, to adverse events.

The I-SPY2 trial was an example of a platform trial in breast cancer where patients were stratified using biomarkers (oestrogen receptor, progesterone receptor, HER2 status, microarray results) and randomised to receive either: paclitaxel with one of three new drugs, or paclitaxel and trastuzumab with one of three new drugs as neoadjuvants prior to surgery.⁹² I-SPY2 led to six new investigational treatments being advanced to phase three trials with each drug matched to biomarker signatures where they were most efficacious.⁹³

Similar approaches are now being pursued in intensive care medicine and infectious diseases. The PREPARE research network is an EU funded platform collaboration which is designed to offer a rapid clinical research-based response to new or re-emergent epidemics.⁹⁴ The MERMAIDS-ARI study is a PREPARE funded, observational platform study of acute respiratory infections in 2000 adults. It was due to finish recruitment in April 2019 but has since been adapted to include patients with COVID-19. The International Severe Acute Respiratory Infection Consortium Coronavirus Clinical Characterisation Consortium (ISARIC4C) is study of COVID-19 that implements a coordinated research response across the UK. Its aims are to phenotype patients using clinical and biological variables using a set of pre-defined and tested research tools created in preparation for respiratory infection outbreaks.⁹⁵

1.6.3 Sepsis

Two recent studies have used whole blood RNA sequencing to identify endotypes in patients with sepsis, one from MARS Consortium and the other from the Genomic Advances in Sepsis (GAinS) study.^{96–98}

The MARS consortium, from the Netherlands, used consensus clustering to find the best clustering method and random forests to select the best genes that classified each endotype. Using the 140 gene-based classifier, the investigators found four groups which they labelled Mars 1-4. 306 patients were used in the discovery cohort, and their findings were validated externally using another cohort from the Netherlands (n = 206) and results from the GAinS study (n = 265, in this paper).

The Mars-1 endotype was found to have the worst 28-day survival and was the most consistent endotype across the validation cohorts in terms of mortality. Using combinations of the

identified genes in the 140-gene classifier and gene expression ratios, they attributed two top-performing genes to each endotype. In Mars-1 these were bisphosphoglycerate mutase (BPGM) and transporter 2, ATP binding cassette subfamily B member (TAP2). BPGM is a 2,3-diphosphoglycerate, which modulates oxygen affinity to haemoglobin. TAP2 is a member of the superfamily of ATP-binding cassette transporters involved in antigen presentation. Enrichment analysis of differentially expressed gene signatures in Mars-1 identified downregulated pathways associated with innate and adaptive cell functions, and upregulation of pathways associated with haem biosynthesis and aberrant metabolic function. The authors suggested the metabolic dysfunction, which had previously been described in sepsis, associated with this endotype represented a failure of immuno-metabolic circuits leading to immunoparalysis and poor survival.⁹⁶

The GAINs study recruited patients admitted to ICU with sepsis due to either community-acquired pneumonia or faeculent peritonitis. Transcriptomic analysis of blood from 265 patients identified a sepsis response signature (SRS) associated with higher mortality and T-cell exhaustion. The authors considered patients with this SRS-1 phenotype to be immuno-incompetent. The SRS phenotypes were discovered using a combination of hierarchical clustering on the 10% most variable gene probes and sparse generalized linear models fitted to mortality. Enrichment and pathway analysis of top 3,080 differentially expressed genes showed functional differences related to T-cell activation, apoptosis, phagocyte movement, endotoxin tolerance and hypoxic response. 41% of the patients were categorised as SRS-1. A seven gene subset was identified as being predictive of SRS-1, which was successfully validated in a cohort of 106 patients. Similar outcomes of organ failure and mortality were observed in the validation SRS-1 group.⁹⁷

The authors proceeded to investigate genomic-level modulation of sepsis by using their gene expression results as a quantitative trait for *cis*- and *trans*-eQTL mapping. These methods enriched for known immune-related pathways and genes (PI3K signalling, antigen presentation, mitochondrial dysfunction). The authors were unable to reproduce an association with the intronic *FER* variant that was described in their previously published GWAS of septic patients.⁹⁹

Of note, was the finding that the Mars-3 and SRS-2 endotypes, both low-risk groups, correlated well with each other. Both were characterised by heightened expression of genes predominantly involved in adaptive immune functions, adding a degree of external validity to both of these studies.⁹⁶

The SRS phenotypes have been incorporated into a secondary analysis of the VANISH randomised controlled trial, which was published in 2016.¹⁰⁰ This study investigated the

use of different vasopressors for cardiovascular support in septic shock (norepinephrine, vasopressin) alongside glucocorticoids (hydrocortisone). The secondary analysis study imputed the SRS phenotypes into the treatment arms *post hoc*, and found that patients with the immunocompetent SRS-2 phenotype had worse outcomes if administered hydrocortisone. This effect was based on small numbers; 31 patients in the SRS-2 hydrocortisone group. This was the first example of how transcriptomic-guided therapy might influence patient outcomes in sepsis.¹⁰¹

Seymour et al. (2019) have recently described four sepsis phenotypes (α , β , γ , δ) derived from a combination of pooled observational studies and randomised controlled trials (PROWESS, ProCESS, ACCESS) of patients with sepsis.¹⁰² The PROWESS, ProCESS and ACCESS studies were all RCTs for patients with sepsis where activated protein-C, goal-directed fluid therapyⁱⁱⁱ and eritoran^{iv} respectively were investigated. Clinical variables were combined with 27 protein biomarkers and the optimum number of phenotypes was derived using a combination of two clustering methods: consensus k-means clustering and ordering points to identify clustering structure (OPTICS).¹⁰³ Genomic data was not included in this analysis. Latent class analysis was used as an independent, confirmatory method to determine the optimal phenotype number, based on Bayesian information criterion and posterior probabilities. The mean values for standardized variables in each group were consistent when comparing the different phenotypes, irrespective of whether the group was derived using a consensus k-means clustering or LCA method.

The α and δ phenotypes were well separated for short term mortality outcomes. The authors suggested that the δ phenotype, which was associated with poor outcomes, cardiovascular and liver dysfunction, aligned with the SRS-1 and Mars-2 endotypes from the GAIN and MARS sepsis studies. Also that the α phenotype, which was associated with better short-term outcomes aligned with the SRS-2 and MARS-2 endotypes.

The authors conducted simulations to determine the outcomes of patients that might have been randomised to different treatment arms of the PROWESS, PROWESS and ACCESS studies. Enrichment of baseline patient characteristics for a given endotype preceded simulation. The expected differences in mortality were compared within the same simulation. The results of these simulations suggested that patients with the more unwell δ phenotype, would have suffered harm from eritoran and goal-directed fluid therapy, whilst the α phenotype would have had better outcomes with goal-directed fluid management. These endotypes were derived from clinical variables and biomarkers without the need for analysis of gene

ⁱⁱⁱ Goal-directed fluid therapy refers to titrated administration of intravenous fluids to a physiological target.

^{iv} An investigational drug for the treatment of severe sepsis.

expression. This study raised the possibility that real-time assignment to an endotype was feasible, and a stratified approach to sepsis management might be available for future studies.

1.7 Successful endotyping approaches in ARDS

Calfee et al. (2014) demonstrated, by using latent class analysis (LCA), that the physiological, biochemical and cytokine characteristics of patients enrolled in the ARDSnet:ARMA and ALVEOLI RCTs, could be combined to define two classes. These two classes were termed hyper-inflammatory and hypo-inflammatory.¹⁰⁴

Latent class analysis is a structural equation model which uses latent structure (hidden groups) to explain outcomes. LCA has been widely used in the social and psychological sciences where it helps to predict behaviours, voting trends and psychopathology. LCA differs from other clustering methods because it estimates a likelihood for the fitted model. Fitted models with different numbers of latent classes can be compared using statistical, likelihood-based methods that are not possible with other unsupervised clustering methods. An optimum number of classes is therefore determined statistically.

The hyper-inflammatory group, who constituted one-third of the enrolled patients, were characterised by higher concentrations of IL-6, soluble TNF receptor-1 (sTNFR-1), plasminogen activator inhibitor-1 (PAI-1), lower bicarbonate and platelets. The degree of respiratory failure did not differ between the groups. This was an important finding since PaO₂-FiO₂ ratio was the established method of differentiating clinical subtypes of ARDS, but it also demonstrated new insights into patients with ARDS.

Stratified treatment responses in each arm of the ALVEOLI study (high PEEP and low PEEP strategies) were identified in patients from each latent class. Patients in the hyper-inflammatory class had significantly improved primary outcome (hospital mortality) if they were randomised to the high PEEP intervention. There were no benefits of a high PEEP strategy for patients with the hypo-inflammatory class.

Follow up studies have applied Calfee et al.'s methods to the FACTT and HARP-2 studies, and they have been able to demonstrate the existence of the same two latent classes in the patients enrolled into both of these trials.^{105,106} Furthermore, following *post hoc* stratification of patients into each treatment arm, they found the primary interventions to have significantly benefited the patients assigned to the hyper-inflammatory class (conservative fluid strategy and simvastatin). The same group have also been able to demonstrate that these endotypes are

stable over time.¹⁰⁷ There is consistency in the methodology used by Calfee's group: using LCA to assign patients in an unsupervised manner, defining classes that do not align pre-existing biases relating to ARDS (degree of respiratory failure, primary diagnosis), external validation in multiple studies. These methods set the results from their work apart from all others to date.

Although LCA lends some insights into ARDS, it is not a complete model. The same group could not replicate their findings in a *post hoc* analysis of the the SAILS study (rosuvastatin in ARDS). Although they successfully identified two latent classes that were consistent with their previous findings, there was no benefit of rosuvastatin to patients classed as hyper-inflammatory. The authors attributed this to the relative lipophilicity of different statins. In addition, one might consider a two-class model as identifying only a single endotype (hyper-inflammatory) whilst assigning the rest to an alternative group. Although the two-class LCA model performed best in each of the studies it was applied to, as determined by Bayesian information criterion, this method did not capture any of the heterogeneity in, or explain any features of the larger, hypo-inflammatory group.¹⁰⁸

Latent class analysis has been used by other authors to describe ARDS endotypes; Reilly et al. (2014) retrospectively studied 1,245 major trauma patients (injury severity score >15) of which 394 developed ARDS (189 derivation, 205 validation).¹⁰⁹ They used LCA to determine three classes which were principally defined by the time, after admission, of developing ARDS. The model was simplified to two groups, using 48 hours as the threshold for defining 'early' or 'late' onset ARDS.

The early group were defined as being more likely to have had thoracic injury, lower blood pressure and received a blood transfusion. This group had higher Ang-2 and sRAGE levels but only the Ang-2 level was significantly higher after correction for multiple comparisons. Mortality was similar in both groups. The authors used a validation cohort and found that thoracic injury and blood pressure were consistent in the early-onset group, but the requirement for blood transfusion was not. Details about model fit, misclassification rates and receiver operating curves were not included.

The authors successfully identified a subset of ARDS due to haemorrhagic shock that had features consistent with the published literature on ARDS and trauma (raised Ang-2). However, their model was based upon incorporating a large number of variables in a relatively small sample size. The frequency of clinical events would, therefore, have been relatively low. This is relevant because variations in data quality and recording during initial acute trauma care compared with late, in-hospital care might influence the model fit. Similar to the other studies listed above, there was minimal exploration of the second identified class.¹⁰⁹

1.8 Background to studies used in this thesis

The results and data collected by three studies, all conducted in the UK, were generously shared to enable the analysis for this thesis to be undertaken. A brief outline of each of these studies follows.

Genomic Advances in Sepsis (GAinS)

The Genomic Advances in Sepsis study was a collaborative observational study of sepsis conducted between 2005 and 2016. The chief investigators were Charles Hinds and Julian Knight. The biological (transcriptomic, soluble immune mediator) characterisation of patient samples has been conducted by the Wellcome Centre for Human Genetics, Oxford, UK. Adult patients with septic shock, admitted to intensive care were recruited by the participating study centres. This study also formed part of the GenOSEPT consortium, an international consortium investigating the genomics of sepsis. 658 patients were recruited in the UK for the GAinS study. There have been a number of publications arising from this study: *FER* gene SNP association with better outcomes in sepsis⁹⁹, sepsis phenotypes SRS1 and SRS2⁹⁷, sepsis transcriptomic responses in faeculent peritonitis and community-acquired pneumonia.⁹⁸

Mechanisms of Severe Acute Influenza Consortium (MOSAIC)

The Mechanisms of Severe Acute Influenza Consortium was established in 2009, following the emergence of pandemic influenza caused by the A(H1N1)pdm2009 virus. The Consortium lead investigator was Peter Openshaw, Centre for Respiratory Infection at Imperial College London. MOSAIC established a network of eleven hospitals in London and Liverpool where hospitalised patients were recruited, along with a network of nine specialist research centres. The Consortium captured the UK's pandemic waves of winter 2009/10 and winter 2010/11. A total of 255 adults and children were recruited to the MOSAIC prospective observational cohort study over these two periods. Healthy adult and paediatric controls were also recruited. Patient recruitment ceased in early 2011. The primary publication arising from this work (Dunning *et al*) was published in 2018.¹¹⁰

HMG-Co-A reductase inhibitors in ARDS-2 (HARP-2)

The HARP-2 study was a randomised controlled trial (ISRCTN 88244364) of simvastatin in ARDS conducted between 2010-14. The rationale for simvastatin use in ARDS was based on a measurable reduction in pulmonary and systemic inflammation in healthy subjects who

were challenged with inhaled lipopolysaccharide (50 μ g LPS) and administered simvastatin compared with those who received placebo.¹¹¹ The chief investigator was Danny McAuley, Northern Ireland Clinical Trials Unit which is part of Queen's University, Belfast. The trial sponsor was Belfast Health and Social Care Trust. 539 patients were randomised as part of this study. The primary outcome was an improvement in ventilator free days (VFDs) by day 28. VFD refers to the number of days a patients is off a ventilator, 28 days after recruitment. This method of outcome measurement will score 0 for both a patient who dies before day 28 and a patient who is still on a ventilator at day 28. The trial did not achieve the primary outcome of a reduction in VFDs by day 28 ($p = 0.06$). The trial results were published in 2015.³⁷

The details of ethical approvals for each of these three studies are available in Appendix A.

1.9 Summary and aims

ARDS came into sharp global focus in 2020 due to the COVID-19 pandemic, yet there have been no direct treatments to date that have addressed the underlying biology in patients with ARDS. The efforts of researchers in this field have been repeatedly hampered by patient heterogeneity and poor characterisation of the biology. Approaches to address this have generally only considered one aspect of the nature of this syndrome - biomarkers, gene expression, physiology, cause. The most successful approach to date by Calfee and colleagues integrated clinical measurements with biomarkers. Integration of the biology of ARDS is required to characterise the underlying mechanisms. The heterogeneity of patients with ARDS mandates that such an exercise will require collation of a large quantity of experimental and clinical data.

The aim of this thesis is to demonstrate endotypes in patients with ARDS using the data collected by these three studies. These endotypes will be characterised by different underlying biological processes and immunological responses in these patients. Demonstration of endotypes may enable future studies to adopt stratified approaches to ARDS and improve the prospects for interventional studies to deliver treatments that benefit patients with ARDS.

1.10 Hypothesis and thesis overview

The hypothesis that this thesis will address is:

“Within the heterogenous clinical syndrome of ARDS there are distinct mechanistic endotypes that can be identified using intergrated phenotyping methodologies.”

The data collected by the GAINs, MOSAIC and HARP-2 studies were shared by their research teams to address this hypothesis. The methods by which the data were integrated and mechanistic insights gained from each of these studies is described in the following chapters.

Chapter 2 describes the methods used to discover sub-types and how these data were integrated to provide insights into the underlying mechanisms.

The results are separated into three chapters: Chapter 3 demonstrates clustering of biomarker and transcriptomic data, and results of applying standard bioinformatic methods. Chapter 4 describes how protein biomarker clusters and gene expression modules were integrated together to define endotypes that were based on biological processes. Finally, Chapter 5 integrates these endotypes with clinical features and outcome data so that they can be described in a patient related context.

Chapter 6 is a general discussion of the findings and possible limitations of this work.

CHAPTER 2

Methods

2.1 Overview of methods

The broad philosophy underlying this project is: “there are subtypes of ARDS that can be identified using biological features and described by their underlying mechanisms. Although these subtypes are defined irrespective of outcomes, they may associate with particular outcomes and clinical features.”

If one considers a single organ syndrome like acute hepatitis, in the absence of patient history, the blood tests that are used to diagnose liver dysfunction (alanine aminotransferase, alkaline phosphatase, bilirubin) may all be grossly deranged due to several causes. The aetiology of acute hepatic dysfunction that causes major derangement of these blood markers may include severe paracetamol toxicity, ischaemic, alcohol-related, autoimmune or viral hepatitis. Trying to model patient outcomes of acute hepatic dysfunction using only these liver function tests without an understanding of the underlying pathology and aetiology could be considered an analogy for the conduct of research into ARDS. This example is not intended to be a criticism of researchers or their methods, but reflects how the processes that might lead to ARDS are far removed from the measures and surrogate markers used to model and predict outcomes for this syndrome. The example also serves, in a crude way, to highlight the differences between diseases and organ failure syndromes in general.

Except for latent class analysis-based methods, the other attempts to endotype ARDS described in the introduction section have used outcomes or clinical variables as dependant variables upon which independent predictors were fitted in regression models. This type of analysis is often referred to as supervised learning, as the regression algorithms attempt to adjust and optimise the influence of independent variables to fit the given labels. The flaw

in these methods is that multiple biological pathways and processes may converge on these outcomes similar to the example of acute liver dysfunction above. Using an outcome label as a predictor, therefore, leads to loss of information from the underlying data structure and only serves to identify features of the final common pathway that occurs during or leading to death, not the initial processes that started this chain of events. For specific diseases with clearly defined mechanisms (for example antibody-driven autoimmunity) well-defined outcomes may better reflect the underlying biology, but given the heterogeneity of patients who develop ARDS, supervised approaches will continue to face these difficulties.

The approach taken for endotype discovery in this thesis has remained unsupervised, with outcomes or other clinical features of the data only incorporated to delineate the differences between subtypes once they already been defined *a priori*.

Unsupervised learning approaches involve the identification of hidden structure or self-similarity between samples. These structures may be referred to as clusters. The number of clusters within a given set of data is therefore unknown. The number of clusters can be derived either using established algorithms or by inspection of the cluster features that are concordant with external information.

If a set of clusters is identified, their features and the method by which these clusters were derived needs to be transparent, as the purpose of this study is to determine the mechanisms that underlie an identified endotype. This requirement precludes the use of neural network-based approaches for unsupervised learning (for example variational autoencoders). Neural networks use multiple layers of nodes, which behave like primitive neurons, with varying weights and activation thresholds to minimise a cost function. They are an effective approach for classification problems and perform with high accuracy. Decoding the features that give rise to clusters using these approaches is, however, non-trivial and may bear no resemblance to the original data after it has been processed through multiple layers of nodes.

It is for the above reasons of methodological transparency that hierarchical clustering and weighted gene co-expression networks were used to determine the endotypes of ARDS for this project. Both of these methods are unsupervised and widely used in the literature for identification clusters and to identify groups of important genes. These methods were also chosen because they are reproducible and do not depend on random starting values (seeds). To determine the differentiating features of each cluster a linear discriminator analysis method was used. The manner in which it transforms data can be visualised easily and the relative importance of contributing features are easily interpreted. An outline of the overall data analysis methodology is shown in Figure 2.1. A detailed discussion of each component of this workflow is presented step-wise in this chapter.

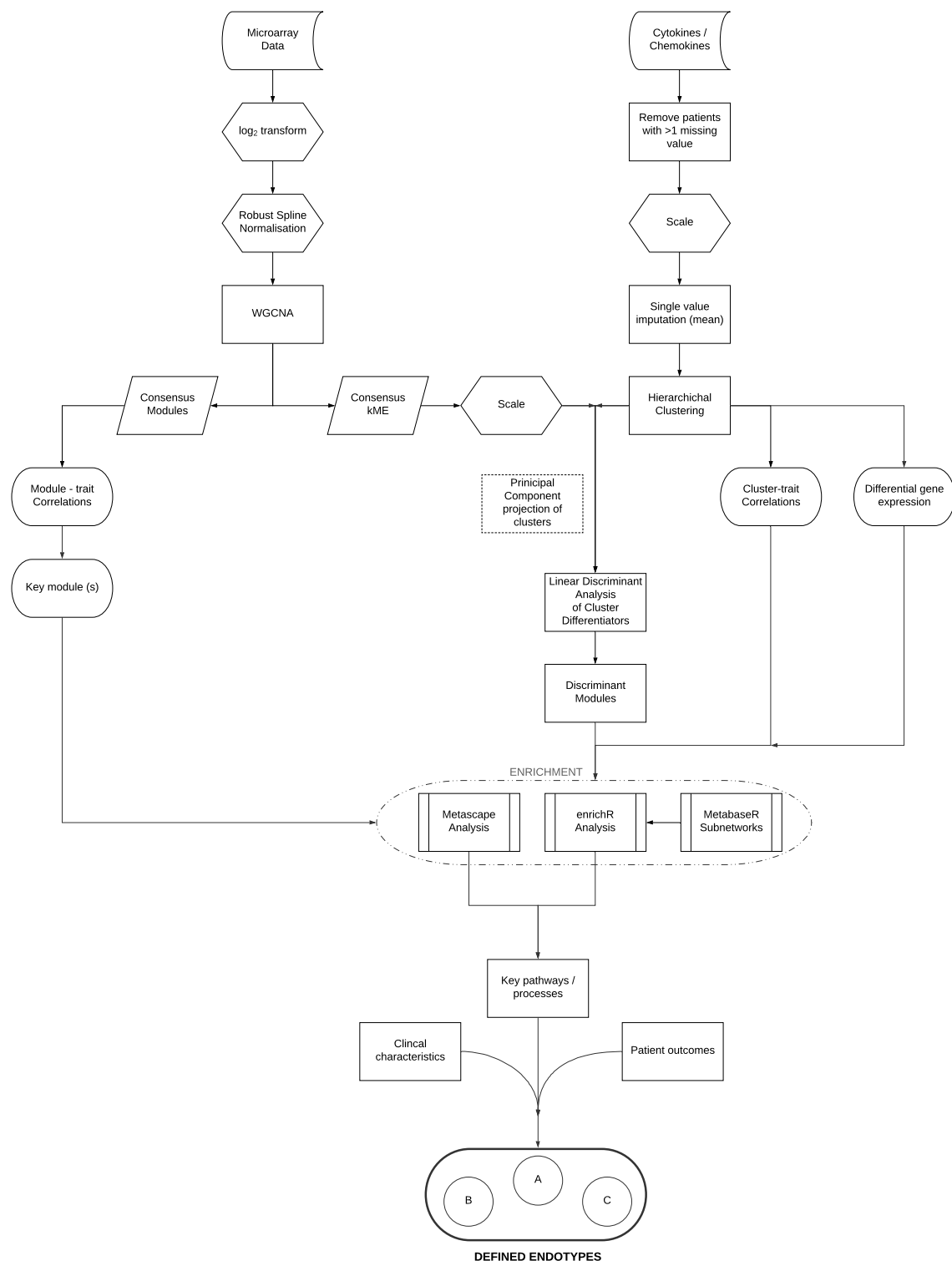


Fig. 2.1 An outline of the methods used in this analysis to define endotypes of ARDS

2.1.1 Data sources

The data for this analysis was generously provided by the research groups who conducted them. The microarray data from the MOSAIC study is publicly available in a Gene Expression Omnibus (GEO) repository (reference:GSE111368). Only subsets of the GAINs data are available in public repositories but are available to collaborators of the Wellcome Centre for Human Genomics, Oxford, UK at request, subject to agreement. The HARP-2 study data is the property of the Northern Ireland Clinical Trials Unit. Sharing of data for this thesis was subject to contract.

A comparative overview of each of these studies is laid out in Table 2.1

	GAinS	MOSAIC	HARP-2
Inclusion diagnosis	Sepsis – Faeculent peritonitis / Community Acquired Pneumonia	Suspected Influenza	ARDS
Exclusion Criteria	Declined / Withdrawn consent <18 years old Advanced directive for withholding life support Palliative trajectory Pregnancy Immune-compromise	Declined / Withdrawn consent	Declined / Withdrawn consent <16 years old Pregnancy Statin contraindication Not intubated / ventilated
Recruitment period	Dec 2005 – Mar 2016	Nov 2009 – Jan 2011	Dec 2010 – Mar 2014
Data collection	Prospective	Retro- and prospective	Prospective
n	658	212	539
Mean Age (sd)	63.4 (15.9)	43.1 (15.2)	53.9 (16.4)
no Male (%)	370 (56%)	110 (47%)	307 (57%)
Intervention	None, observation only	None, observation only	CTIMP: Simvastatin (n = 259)
Control group	Cardiac surgery	Healthy controls	Placebo (n = 281)
No with ARDS	317	Not formally attributed	539
Physiology	Day 1 – Day 7	Day 1 – Day 14	Day 1 – Day 28
Biochemistry / Haematology	Day 1 – Day 7	Day 1 – Day 14	Day 1 – Day 28
Imaging	Chest radiograph report	Chest radiograph features	-
Biomarkers	Multiple cytokines / chemokines at multiple time points	Multiple cytokines / chemokines at multiple time points	5 cytokines / chemokines at multiple time points
DNA	Yes	Yes	Collected, analysis pending
Whole blood	Yes (multiple time points)	Yes (multiple time points)	-
RNA			
Plasma	In progress	-	-
Proteomics			
Treatment	Organ support Antibiotics	Organ support Antibiotics Antivirals	Simvastatin / Placebo Organ support
Outcomes	ICU / Hospital Mortality, Organ failure	Hospital mortality, Severity of respiratory failure	28d / 90d / ICU / hospital / 1yr mortality, Quality of life measures

Table 2.1 Overview of each of the studies used in this project contrasting their differing features and common elements.

2.1.2 Sampling times and patient status

Each of the acutely unwell patients recruited to the contributing studies for this thesis will have had a unique illness trajectory. The time-frame in which a patient developed symptoms, presented to hospital, was admitted to critical care, had a diagnosis compatible with recruitment to a study and underwent biological sampling was subject to considerable variation. If repeated samples were taken then these could have also been at different stages of their acute illness: during deterioration or recovery from their primary illness. This variation in patient course is illustrated in Figure 2.2.

Figure 2.2 serves to demonstrate the difference between scientific experiments conducted in the laboratory with controlled parameters and the heterogeneity of clinical studies involving acutely unwell patients. Sampling patients at set time points, for example, admission to hospital or critical care, are subject to variation in the patients' symptoms, degree of physiological derangement and their interactions with the health systems treating them. Patients may be at different stages of their evolving illness at these predetermined times and thus the biology captured by biomarkers or gene expression will be subject to the same variation.

An additional source of variation occurs for patients recruited after admission to intensive care who were transferred from another intensive care setting. Academic health centre hospitals often receive patients from smaller hospitals for specialist management. These centres, where many patients for critical care research studies are recruited, have local research infrastructure and higher capacity intensive care units. Recruitment to clinical research studies of critically unwell patients may, therefore, be over-represented by patients admitted to academic health centres who may not be representative of the wider population.

Figure 2.2 also illustrates how the outcome of death (red asterisk and red line) may be an additional source of variation. The figure denotes these events as "palliative trajectory" which reflects a change in the focus of care to symptom control and withdrawal of life-sustaining treatment. These decision are made by treating teams where they believe, on the balance or probabilities, that maximal therapy is deemed to have failed, will only prolong the patient's suffering and is unlikely to change the ultimate outcome. Decisions concerning the limitation of critical care interventions are subject to individualised consideration of the patient's illness in the context of their pre-morbid functional status, patient's wishes, patient family's knowledge of a patient's prior expressed wishes (where the patient is incapacitated) and the biases of the treating clinical team.

It could be hypothesised that humans, broadly speaking, only have a finite number of stereotyped responses to severe acute illness, although they may transition between these different states during their illness. Assigning labels to patients based on clinical events and outcomes will miss this heterogeneity as different processes may result in similar outcomes. Using these labels in analytical approaches may therefore direct the analysis incorrectly. Consideration of all sampling time points consistent with acute illness (but not convalescence) might offer the opportunity to learn the features that describe commonly occurring biological states. This requires taking the unsupervised learning approach described above, but also to use all the available biological data from all sampling times. There is a trade-off here as restricting the analysis to recruitment day samples alone might produce results that can be applied to clinical practice. The data collected at study recruitment is more likely to reflect the patient populations recruited to observational studies and clinical trials. Using multiple sampling times requires many variables to ensure identified states are robust, otherwise small deviations in a single variable between different sampling times will cause instability of identified states.

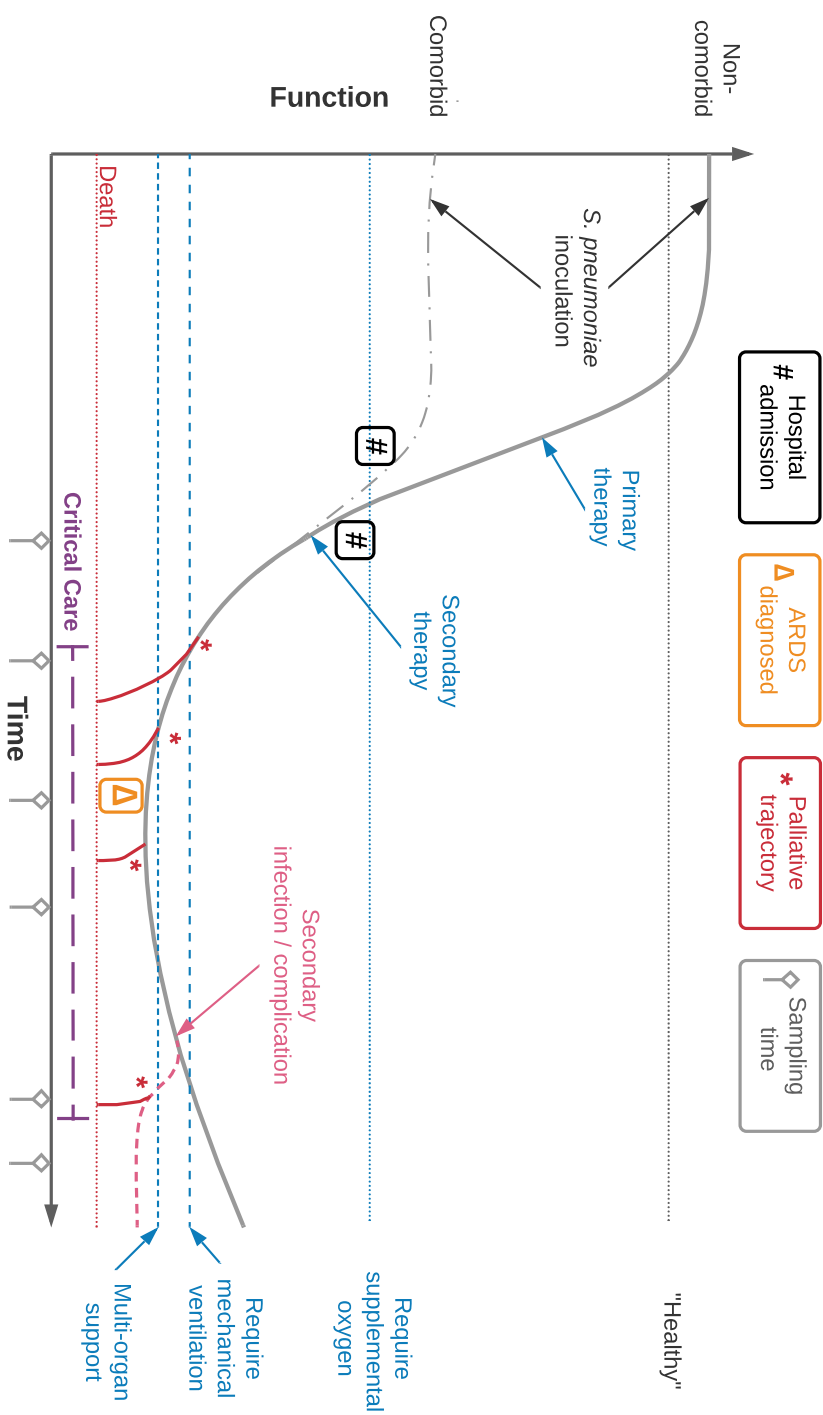


Fig. 2.2 Hypothetical illness trajectories for patients, with or without co-morbid conditions, who develop an acute illness (*S. pneumoniae* pneumonia) and are considered for admission to critical care and subsequent recruitment to an observation study or clinical trial. Every stage of illness, diagnosis and intervention is subject to individual variation and so the potential sampling times (grey markers on time axis) at which data is recorded and biological samples taken from patients is likely to reflect a different set of biological processes and states for each patient sample in a study. The trajectory for patients' recovery from acute illness is not considered in this diagram. Primary and secondary therapy refers to interventions in primary and secondary care settings. 'Healthy' refers to an arbitrary functional state for a given patient but might be represented by Rockwood clinical frailty score ≤ 2 .¹¹² "Palliative trajectory" refers to withdrawal of life-sustaining treatment or death despite maximal therapy *S. pneumoniae*: *Streptococcus pneumoniae*

2.1.3 Biological sample collection, processing and analysis

This project used the results obtained by analysis of clinical samples from three different studies:

1. Genomic advances in sepsis (GAinS)
2. Mechanisms of severe acute influenza (MOSAIC)
3. Hydroxymethylglutaryl-CoA reductase inhibition with simvastatin in Acute lung injury to Reduce Pulmonary dysfunction. (HARP-2)

The background for each of these was described in section 1.8. The research teams that were responsible for each of these studies collected, processed and conducted the experiments that generated these results. The methods for how each study conducted these steps are reported in primary publications associated with them. Their laboratory methods have been included here in Appendices B and C.

2.1.4 ARDS diagnosis

To ensure that comparisons between patients recruited to each study were consistent, the Berlin criteria for diagnosis of ARDS (Section 1.1) was used to identify ARDS cases.⁴ Many of the patients recruited to the GAinS study preceded publication of the Berlin definition. For patients recruited to the GAinS study, the Berlin criteria was applied using the available arterial blood gas, ventilator (PEEP setting) and radiographic features that were shared by the GAinS research team.

The inclusion criteria for recruitment to the HARP-2 study stipulated a diagnosis of ARDS as per the Berlin definition.

Patients recruited to the MOSAIC study preceded the Berlin definition. There was incomplete radiographic and mechanical ventilator data in the study database and so it could not be determined which patients could be diagnosed with ARDS retrospectively. The study authors used the respiratory SOFA score to grade the severity of respiratory failure. The components of this score are listed in Table 2.2.

The focus of this thesis was to determine the endotypes of patients with critical illness. Patients recruited to the MOSAIC study were in the midst of a pandemic and so patients with respiratory SOFA score equal to one were excluded for this analysis. This patient cohort was considered to be too heterogeneous as a positive test for influenza infection in these patients may have been in the context of another acute illness requiring admission to hospital.

Patients with respiratory SOFA scores greater than equal to two from the MOSAIC study were considered for analysis in this thesis. The biological data from these patients was more likely to capture the relative differences between patients with moderate and severe respiratory failure attributable to influenza infection alone. Patients with a respiratory SOFA score equal to one were excluded from analysis.

PaO₂-FiO₂ ratio, mmHg (kPa)	SOFA score
≥ 400 (53.3)	0
< 400 (53.3)	1
< 300 (40)	2
< 200 (26.7) and mechanically ventilated	3
< 100 (13.3) and mechanically ventilated	4

Table 2.2 Respiratory component of the sequential organ failure (SOFA) score

2.2 Hierarchical clustering

Hierarchical clustering is a method that separates data into groups that are distinct from each other and are broadly similar within groups. Projection of data points to a multi-dimensional space is based on their numerical properties and number of features, with each feature adding an extra dimension. The distance between objects can be measured in this projected space based on their coordinates or location. Clustering methods can be classified as being agglomerative (start with individual data points and group neighbours together) or divisive (start with all data points as a collection and divide them step-wise until you arrive at individual data points).

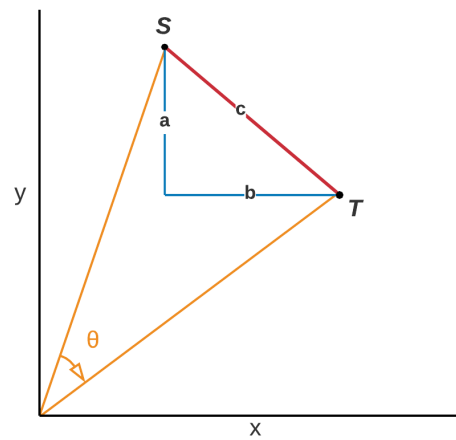
The distances between data points can be calculated in the ways shown in figure 2.3. Both the Euclidean and Manhattan methods are special cases of the Minkowski method (equation 2.1) which describes distances (D) in n - dimensional space, where $p = 1$ for Manhattan and $p = 2$ for Euclidean distances. Cosine similarity is used to compare the similarity between vectors in high dimensional space and plays a role in text mining and computer vision.

$$D(X,Y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (2.1)$$

In biological data, the Euclidean distance method tends to be preferred as this measure incorporates all the available information about each data point. Data should be scaled before calculation of Euclidean distances or else the influence of points will be determined by their magnitude instead of whether they are outlying with respect to their distribution. This is particularly important with biological assays, some of which have a large dynamic range whilst others can only measure quantities of an analyte within a small range of values.

Clusters are determined by the distance between groups of neighbouring objects. This is performed based on a number of predetermined algorithms that determine how one cluster is linked to the next (referred to as linkage). The different linkage methods are represented in figure 2.4.

In order to minimise the amount of variation between members of a cluster the Ward method of clustering was used in the this project. These methods are part of the standard R software statistical package (version 3.6.2). The Ward linkage method used was called by the argument “Ward.D2” in the *hclust* function. This function squares the distances before calculating within cluster sum of squares. The other linkage methods were also assessed prior to deciding upon the Ward linkage as the preferred method.

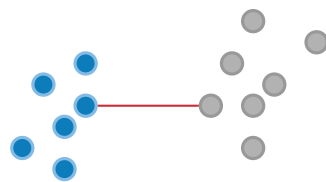


Euclidean distance = $c = \sqrt{a^2 + b^2}$

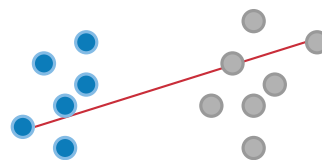
Manhattan distance = $a + b$

Cosine similarity = $\cos \theta = \frac{S \cdot T}{\|S\| \|T\|}$

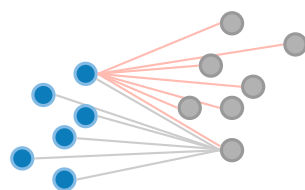
Fig. 2.3 Distance methods used in hierarchical clustering.
 $S \cdot T$ represents the dot (scalar) product of the vectors S and T



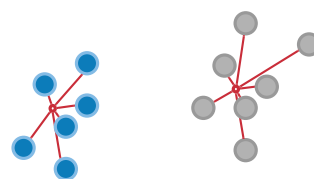
Single linkage:
smallest possible distance



Complete linkage:
greatest possible distance



Average linkage:
average of all distances
between all points



Ward linkage:
minimisation of within cluster
sum of squares error

Fig. 2.4 Linkage methods used in hierarchical clustering

After the distances and linkage methods between individual points have been calculated, pairs of points that are similar are grouped together. The algorithms for distance and linkage are then repeated, using the new groups instead of individual. Each step produces a new join between groups, called branches, containing more members than the previous until all points have been joined. The distance between points and how they join can be represented in a diagram called a dendrogram, which resembles an inverted tree.

The dendrogram linking all data points together may be cut to produce groups of data points, called clusters. The cut height will determine the number of clusters and assign a membership label to each terminal leaf of the dendrogram (Figure 2.5).

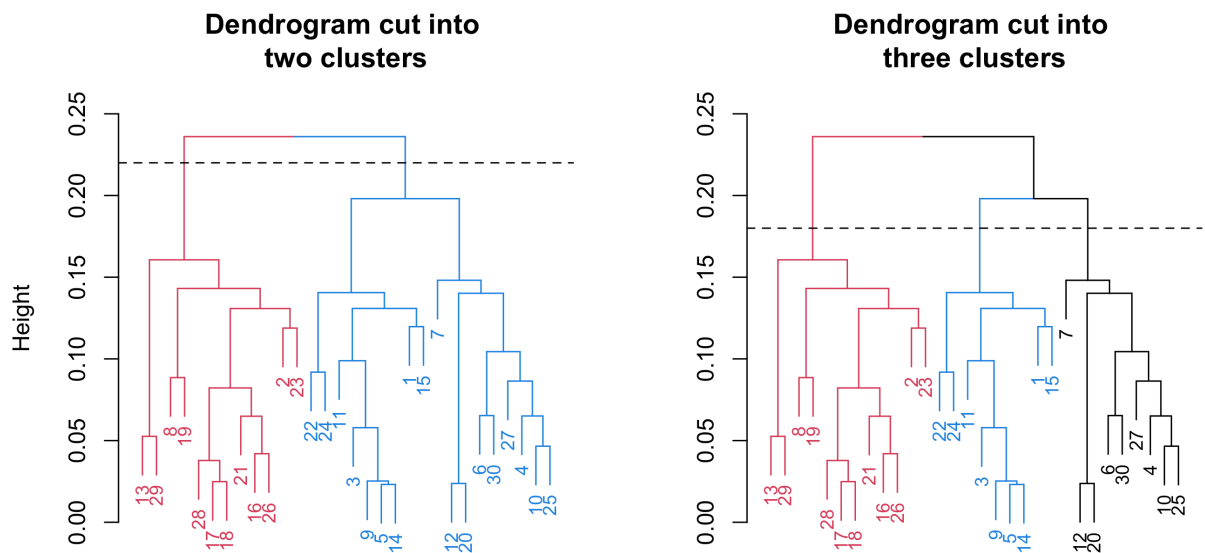


Fig. 2.5 Demonstration of how a dendrogram can be 'cut' (dashed line) at different heights to produce distinct clusters. Colours represent the different branches assigned to each cluster.

There is no standardised method for deciding where to cut a given dendrogram tree. With two or three dimensional data, visual inspection allows for clusters to be observed directly. Once data exceeds this then visual methods are not possible without intermediate steps that reduce dimensionality, but these may remove information and fail to observe adjacency in non-visualised dimensions. One of many possible methods to determine the optimum cluster number is to use an alternative clustering method. For this thesis, the alternative clustering method used was k-mean clustering.

K-means clustering is a semi-supervised clustering method where a given (k) number of centres (centroids) are randomly placed in the data space. The distance from each data point to a given centroid is calculated and the centroid position is moved so as to minimise the within cluster sum of squares error (WCSS) value. The algorithm then iterates this process until all the centroids are stationary at each iteration. The total WCSS can then be plotted per number of centroids and this produces a characteristic scree or 'elbow' plot where the WCSS decays exponentially. The point of maximum curvature prior to flattening of this line, usually determined by inspection, is conventionally considered the optimum number of clusters (k -centroids) as it represents a trade-off between over-fitting and classification error (Figure 2.6).

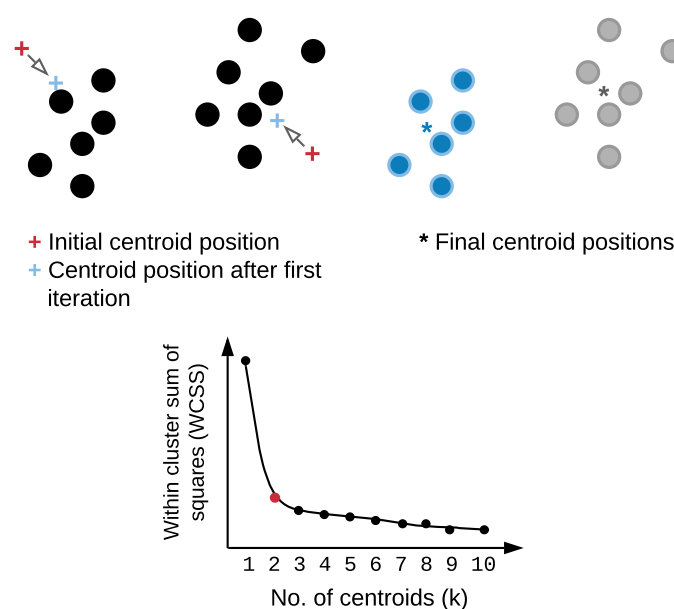


Fig. 2.6 Schematic of how k-means centroids migrate to find their stable locations. The graph is a representation of a scree plot showing the elbow method of identifying (by inspection) the optimum number of k (highlighted red in this case).

K-means is referred to as a semi-supervised method as the number of centres has to be stipulated by the observer of the data, and is not automated. Convergence to stable centroids may also depend on initial starting conditions as the initial points are random. It is not always possible to determine the elbow of the scree plot if the data is unstructured and in these cases the WCSS may decrease linearly with k instead of exponentially.

The elbow method favours clusters that are circular or spherical shaped and it is these patterns of data which will give the highest scores and sharpest elbow plots. There are also many other

indices that estimate the optimal number of clusters. For this project the *Nbclust* package in R, which calculates thirty different clustering metric measurements, was used to determine the optimum number of clusters by majority vote.¹¹³

2.2.2 Management of missing data

Experimental failures and erroneous results can lead to missing values in biological data. The experiments for measurement of protein biomarkers were carried out in other institutions and the details of reasons for missing values or experimental discrepancies were unavailable. It was assumed that the results from these experiments had undergone adequate quality control as part of any standard ELISA protocol. Nevertheless, there were some missing data points in the protein biomarker results from both the GAINs and MOSAIC studies. The management strategy for missing values in data analysis can involve exclusion or imputation of the missing value. Samples with a high proportion of missing values should normally be excluded whilst single missing values are usually imputed. There is no set threshold for when to exclude data or impute data points and this subject to the investigators own understanding of the data, its provenance and the implications of either imputation or exclusion.

If missing data is to be imputed there are many possible ways of achieving this. The simplest method is to use the mean value for the assay for all other samples. This method reduces the variance of the data. A popular method in the data science literature is to use multivariate imputation by chained equations (MICE), which is available via the *MICE* library in R.¹¹⁴

The *MICE* library uses conditional multiple imputation with ordinary least squares (OLS) regression to impute missing values for each variable. This is repeated several times to produce multiple imputed data sets. A further OLS regression addresses the uncertainty associated with these imputed values and a pooled repeated analysis dataset is created. The final regression coefficient for the variable to be imputed is the mean of these pooled coefficients. Imputation using *MICE*-based techniques perform well and are widely used in the literature.¹¹⁵ Potential limitations of the *MICE* method include the need for a random seed as some steps involve random numbers. This has implications for ensuring reproducible results.

Both mean and *MICE*-based imputation were assessed in this project for the protein biomarker data. Cluster assignments based on the complete (no missing values) data were compared with cluster assignment following imputation using either of these two strategies. Cluster concordance was assessed using the Rand index which measures concordance between different clustering methods. The Rand index is bound by the values 0 (no concordance) and 1 (complete concordance). Unlike simple accuracy calculations, the Rand index can take into account different cluster labels and the arrangement of clusters. The adjusted Rand index (ARI) is a more conservative application of the Rand index which involves corrections for

grouping of the data if they had arisen by chance.¹¹⁶ The correction methodology applied in ARI overcomes the tendency for randomly generated data to have a high Rand index as the number of samples increases in a data set. The imputation method with the highest adjusted Rand index score was used.

2.2.3 Cluster stability

Unlike regression-based methods a hierarchical cluster model does not have an associated estimation of likelihood or cost when the model is fitted to the data. This is because all values are assigned to a cluster so there is no way to determine an error or residual values. Although the partitions are consistent within the same data, there is no guarantee this will be the case with unseen (hold out) data, even if this new data is derived from the same parent distributions. To assess stability and whether clusters were reproducible, the data was split into fractions (70:30) and each fraction was clustered independently. Cluster assignments from each split were compared with cluster assignments using all the data by calculating the adjusted Rand index. This process was bootstrapped by re-sampling 500 times, using the same splitting ratio. The mean Rand index with confidence intervals were reported. This method only accounted for cluster partitions and their relative sizes, not the properties associated with data contained in each cluster.

2.3 Microarrays

DNA microarrays are an early implementation of high throughput gene expression analysis. A microarray consists of a chip with DNA probes of known sequence attached that complement segments of known gene sequences. Microarray experiments require mRNA extracted from cells to be converted to complementary DNA (cDNA) using a reverse transcriptase (RT) enzyme reaction. The cDNA is then amplified using a polymerase chain reaction (PCR), labelled with fluorescent dyes and placed on the microarray plate. cDNA sequences that are complementary to probe sequences will form hydrogen bonds and hybridize, remaining attached, whilst sequences that are not complementary are washed off. The reaction between probes and cDNA depends on the chemical and environmental conditions present at the time of the experiment. The fluorescent dyes, attached to hybridized cDNA, are excited using a dual channel laser. The intensity of the signal is detected by a scanner and translated into a numeric value which represents the relative expression of the gene compared with a control sample. (Figure 2.7).

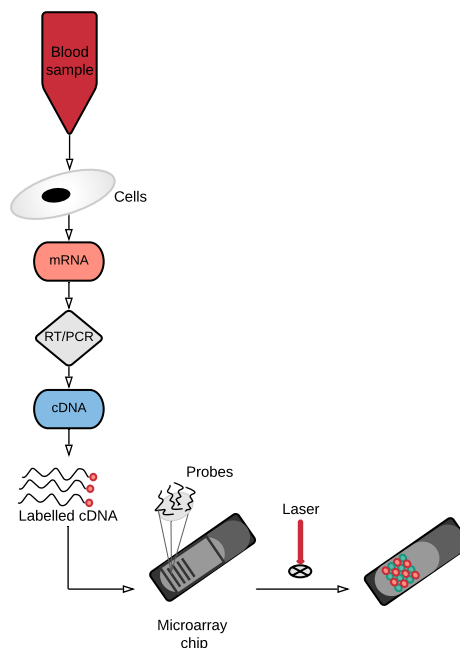


Fig. 2.7 Schematic of a microarray experiment.
RT/PCR: Reverse transcriptase, polymerase chain reaction

The probe sequences, chemistry of how they hybridize and version of the target organism genome used to synthesise the probe sequences vary by chip versions and manufacturer. Fortunately for this project, the GAINs and MOSAIC studies, by coincidence, both used the same microarray version and manufacturer (Illumina HumanHT-12 v4 BeadChip) to quantify gene expression. However, the representation of genes from one microarray chip to the next can still vary due to the manufacturing process, even if chips are the same version. As the probe sequences are fixed, they may fail to hybridize with variant sequences that contain mutations and so the expression of these genes will not be captured by these experiments. cDNA sequences may also hybridize with probes if they have partially complementary sequences. This is known as cross-hybridization and can be a source of false-positive results.

2.3.1 Preprocessing and management of batch effects

Results from microarray experiments therefore require careful quality control and review before inferences can be made. Standard processing workflows for Illumina microarray data are described in the documentation for two R packages *lumi* and *limma*.^{117,118}

The probe intensity results from a microarray experiment should, in theory, form a normal distribution. In practice, many probes have intensities with high variance and there is a significant amount of experimental noise, which is more apparent in low-intensity probes.¹¹⁹ The raw intensity data tends to display a long right-sided tail and so intensity values are initially log-transformed.

The second step is quantile normalisation: the relative intensities of each probe are ranked across the genes, an average rank is calculated across all samples and intensity values are replaced by these new averaged ranks. This preserves the ranks of the genes but information is lost with respect to their relative intensities. A modification of this method is ‘robust spline normalisation’ (RSN) which combines quantile normalisation with a continuous transformation using a regression spline fit. The parameters for this spline fit are estimated by comparing probes that are strongly differentially expressed. An estimated intensity for each probe is then calculated using this estimated parameter as a scaling factor. There are many methods for normalising microarray results; RSN performs well when compared with alternatives.¹²⁰

After normalisation, probe intensities often vary between microarray experiments that have been carried out at different times, even if they are chip same version. Sources of this variation can occur at any point in the pathway shown in Figure 2.7. Ambient environmental conditions can influence enzyme performance and efficiency of amplified cDNA hybridization with the array probes. These differences attributed to external conditions that are unrelated to the biology being tested in the experiment are often referred to as ‘batch effects.’

There are a variety of methods to adjust for batch effects. A widely used method that is easy to implement is the ‘ComBat’ method from the *SVA* library in R.¹²¹ This function uses an empirical Bayes transformation where the background gene levels in each batch are used to estimate the prior distribution. Variances are pooled across arrays to ‘shrink’ each batch.¹²² The overall effect is to shrink the mean and variance of the expression levels of the genes. Compared with other batch effect correction methods, ComBat performs well and is widely used.¹²³

2.4 Weighted gene co-expression network analysis

Weighted Gene Co-expression Network Analysis (WGCNA), recently renamed weighted correlation network analysis, uses network-based methods to resolve the challenges of interpreting data characterised by repeated sample measurements using large numbers of features. A practical example is the results from microarray experiments where several thousand probes are measured simultaneously from samples subjected to different experimental conditions. As the number of probes dwarfs the sample size, meaningful comparisons are difficult to demonstrate reliably after multiple comparison correction. Genuine signals in these data may be statistically suppressed without a large sample size with low experimental heterogeneity.

2.4.1 Network construction and intuition

WGCNA, published in 2005, could be described as inferring guilt-by-association, where the strength of the correlation determines the association. The first step is to calculate the correlations between all measured values within a sample (e.g. gene probe intensities), across all measured samples.¹²⁴ This is performed using Pearson's correlation across all probe pairs which produces a similarity matrix. If two gene probes are expressed at similar levels between samples consistently, then they will have a consistently high correlation. In graph and network theory, the correlations are interpreted as the weights of edges between each of the nodes. The correlation matrix can be interpreted as a weighted network (Figure 2.8).

Simplification of this network uses thresholds which prune edges and identify consistently linked communities of nodes. A hard-threshold approach refers to the pruning of connections that are below a set value ($r < 0.8$ in Figure 2.8). Unweighted networks are generated by a hard-threshold approach and are useful for observation of the network interactions within a focused set of biological processes. If the focus of the analysis is to determine a global view of the network, which incorporates contributions from all processes, then unweighted networks are preferred.

Weighted networks may preserve information between nodes but are noisy and are described as random networks. In random networks, the number of connections between nodes tends to form a normal distribution and are uninformative. Biological networks are generally considered examples of scale-free networks where key genes or proteins control the function and expression of many others. A scale-free network is one where the probability for a given node to be connected to another decays as a power law or shows a power law-like asymptotic

distribution (Equation 2.2). In a scale-free network, most nodes have very few connections and a few have many. This property persists independent of network size. Scale-free network phenomena can be observed when studying the links between websites on the internet or interactions on social media platforms. Several key nodes have concentrated connectivity (e.g. google.com). New nodes added to the network are more attracted to these established nodes (preferential attachment), which are referred to as hubs.

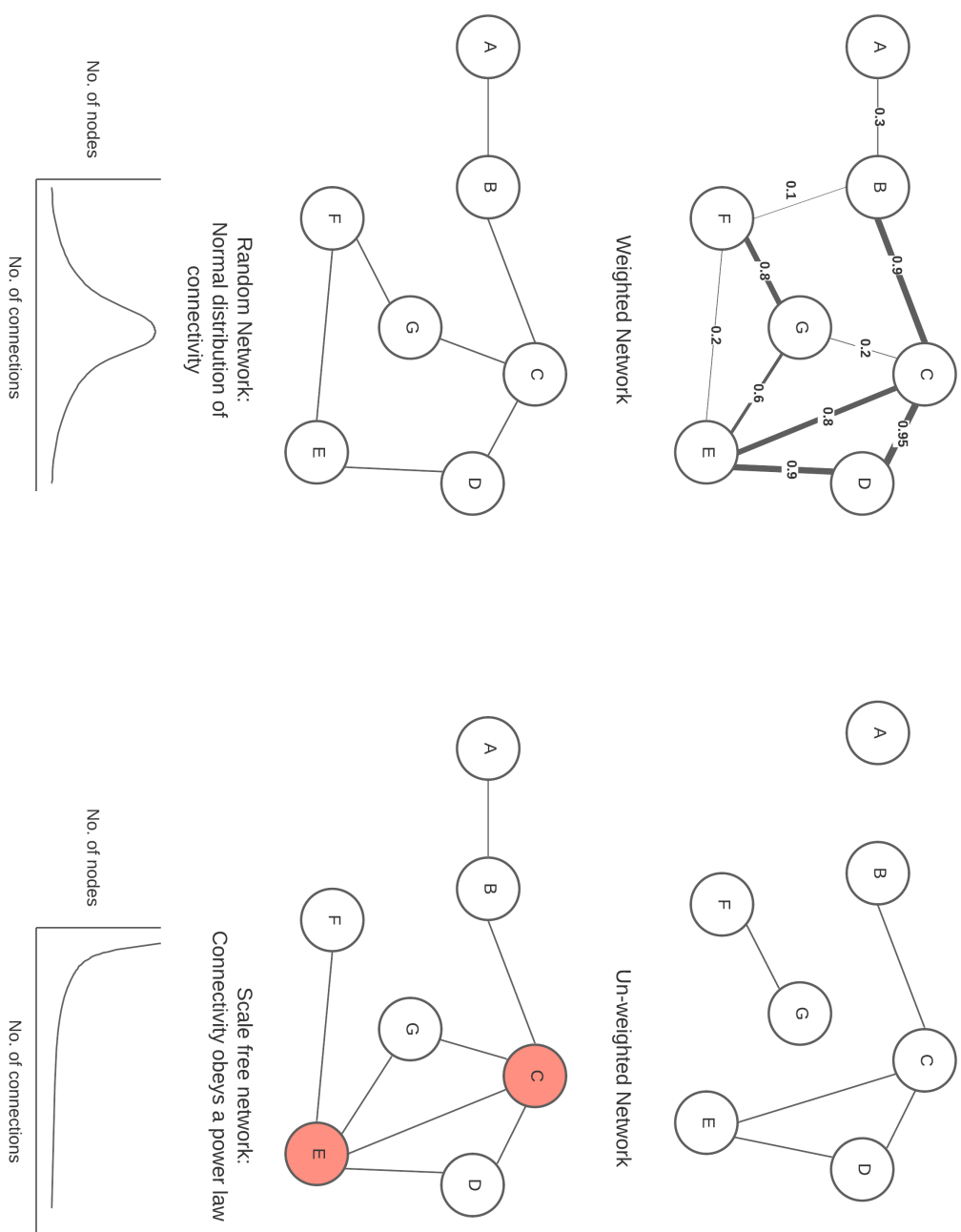


Fig. 2.8 Representation of graph theory concepts applied to gene expression using seven example genes / nodes (labelled A-G). In a weighted network, information relating to the magnitude of all edges (calculated using Pearson's r) is preserved. In un-weighted networks, connections not achieving a high enough threshold (in this example $r < 0.8$) are pruned. Scale-free networks obey a power law where key nodes (highlighted in red) are more dominant over the others in terms of their number of connections. These are referred to as hubs.

To translate a weighted-network into a scale-free network a soft-power threshold is used. The connection weights are raised to an increasing power value (β , Equation 2.3) until the connectivity distribution between nodes resembles one described by the power law.

The power law:

$$P(k) \sim k^{-\gamma} \quad (2.2)$$

k is connectivity, $P(k)$ is the probability of a node being connected
 γ is a constant calculated as the slope of the $\log P(k)$, $\log k$ plot.
 Generally, $3 \geq \gamma \geq 2$

Adjacency using a soft power threshold:

$$a_{ij} = |\text{cor}(x_i, x_j)|^\beta \quad (2.3)$$

a is the adjacency matrix of i and j
 x_i is the expression of i

The generalized topological overlap-based dissimilarity measure:

$$d_{ij}^{T,[m]} = 1 - t_{ij}^{[m]} \quad (2.4)$$

d_{ij} is the dissimilarity of genes (nodes) i and j ,
 $T, [m]$ is the m -th order TOM
 $t_{ij}^{[m]}$ is a generalised TOM

Fig. 2.9 Equations for network analysis of gene expression used in the WGCNA library. TOM: Topological overlap measure

Larger values of β will reduce adjacency, causing a greater degree of pruning and the resulting networks will appear to be more sparse. To arrive at an optimum value of β the, r^2 of the linear model between $\log_{10}P(k)$ and $\log_{10}k$ is calculated. $r^2 > 0.8$ is considered to be sufficient to represent a scale-free network that obeys the power law. A scale-free network allows for measurement of overlap between groups of connected genes and all the other genes in the network. This is referred to as topological overlap measure (TOM). TOM implementation in WGCNA allows for recognition of connections between higher order neighbourhoods (communities that are slightly further away).

A dissimilarity measure is calculated on a higher-order TOM (Equation 2.4). Hierarchical clustering with average linkage produces dendrogram branches that are more pronounced (separated by larger distances) than other dissimilarity methods.¹²⁵ In contrast to standard dendrogram cutting of hierarchical clusters, the WGCNA authors developed a novel branch cutting method called ‘dynamic tree cut’. This approach uses a flexible dendrogram cutting method, where the minimum module size and minimum cut height can be directly stipulated, to aid detection of groups of connected genes.¹²⁶

The result of dynamic tree cut is the isolation of dendrogram branches into groups of coexpressed genes, referred to as gene modules.

2.4.2 Downstream analysis of results after applying WGCNA

Genes within a module can be said to be highly connected and have a similar co-expression profile. Each gene module is assigned a module eigengene (ME) property, a representative gene for each module. An ME can be considered the first principal component of a given module. The relative distances between MEs were calculated and clustered on a dendrogram to visualise how closely related individual modules are to each other. Unassigned genes were labelled as ME0 (‘grey’) and these represented the group of poorly connected, presumed to be background, genes.

Data from microarray and other high throughput genomic experiments are subject to the $p \gg n$ problem where experimental results generate many more variables than the number of available samples. In addition, results from high throughput experiments are often noisy. If standard regression algorithms are used to make inferences, these models may over-fit and perform poorly with new, unseen data. WGCNA manages this problem by identifying genes that are consistently poorly connected to others in the network. Unassigned genes can either be treated as a single entity or isolated from downstream analysis. The remaining modules may contain as few as fifty genes or many thousands. WGCNA, therefore, effectively reduces the noise from high dimensional data and facilitates meaningful analysis of gene expression data by organising genes into highly connected groups. The module eigengene concept has additional utility; a large number of genes can be represented by a single value that can be used for correlation with experimental conditions or sample traits.

WGCNA has been used successfully to explore the differentially expressed genes in muscle tissue from patients with ICU-acquired weakness.¹²⁷ The investigators correlated gene modules with patient phenotypes (muscle strength and function) at different stages of their recovery (seven days and six months after discharge). Gene module function was determined

using enrichment analysis of the constituent probes. Pathways from the key modules were associated with mitochondrial function, calcium handling, muscle structure development and extracellular membrane deposition (healing and repair). The authors validated their findings in an experimental porcine model and independent ICU cohort.

The WGCNA library contains a *consensusBlockwise* function which can compare samples from multiple microarray experiments simultaneously, as long as they have a sufficient number of probes in common. The GAINs researchers conducted gene expression quantification for recruited patients with four separate microarray experiments. Each microarray result contained data from a mixture of three groups of patients (community-acquired pneumonia, faeculent peritonitis and cardiac surgery). Before using this function, the soft power threshold values for each of the microarray experiments were determined, to ensure that the networks derived from each microarray experiment had similar properties.

2.5 Data integration and linear discriminant analysis

Correlation coefficients between gene modules clinical variables were calculated to establish whether modules were associated with clinical phenotypes. p values were corrected using the Benjamini-Hochberg method.

The protein biomarker values from each contributing study were segmented into plausible clusters based on consensus hierarchical clustering using the *NbClust* library.¹¹³ The differences between identified clusters were explored using linear plots of averaged, grouped values for each cytokine and with heatmaps. These relationship between points in assigned clusters were visualised using principal component analysis (PCA). PCA is a dimension reduction method for high-dimensional data which maximises the variance of data along fitted orthogonal axes, using singular value decomposition (SVD). It can, therefore, capture the dominant directions of variance in the data. SVD is an algebraic manipulation of matrices that is widely used in many statistical and data analysis methods. Each principal component explains the data variance in sequentially decreasing proportions.

Gene modules identified by WGCNA have several numeric properties. For each sample, the algorithm calculates the explained variance for the relationship between the sample and the first principle component of each identified gene module. The package refers to these values as ‘kME’s. kME values were zero-scaled on the mean to enable integration with the protein biomarker data. Clusters were re-visualised in the principal component space to determine whether gene expression results from WGCNA disrupted the structure or arrangement of points in each cluster. The relative loadings of cytokines and kMEs values

were visualised in a ‘biplot’. A biplot is a projection of data points using the values from the first two principal components as horizontal and vertical axes. Arrows are drawn for each variable on this plot from the (0,0) co-ordinate. The direction and magnitude of each arrow represents the explained variance (loading) of that variable with respect to the first two principal components. Arrows that point horizontally contribute to the explained variance of the first principal component, vertical arrows to the explained variance of the second principal component.

2.5.1 Linear discriminant analysis

A linear discriminant analysis (LDA) model was fitted to determine the most discriminant features between clusters. LDA is similar to PCA in that it defines new linear functions (axes) to fit the data. The key difference between LDA and PCA is that LDA is supervised and seeks to maximise the differences between assigned labels. PCA, on the other hand, is unsupervised and it defines m new axes that account for the variance along orthogonal axes, where m is the number of variables in the data.

Both the PCA and LDA calculate the relative loadings of contributing variables for their fitted axes. The magnitude of a loading represents its contribution to the principal component in question. In LDA only $k - 1$ discriminant axes are fitted, where k is the number of prior groups. These axes can also be considered decision boundaries. The loadings in LDA represent the relative contributions of the variables to the discriminant axes. To calculate these loadings LDA uses a comparison matrix of cross-products and within-group sum of squares, whereas PCA uses a similarity matrix.

Figure 2.10 demonstrates principal component and linear discriminant analysis approaches for Fisher’s iris flower data. There are four measurements for each of 150 flowers that belong to three different species. The PCA biplot (**A**) shows the data projected on to the first two principal components. The red arrows demonstrate how each of the variables contributes to the respective axes. Here petal width and petal length primarily contribute to the first principal component (PC1); the loadings are parallel to the PC1 axis. Sepal width contributes to the second principal component (PC2) as its loading is orientated in the PC2 direction.

B is a representation of the linear discriminant decision boundaries if they are projected into the PC1 and PC2 subspace. LD1 and LD2 are both almost orthogonal to PC1 and parallel to PC2. If the data points are then projected, using matrix multiplication, onto the LD1 and LD2 axes then we see them distributed as shown in **C**. The groups appear better separated.

There is less overlap between the *versicolor* and *virginica* species when they are projected on to the LD1 axis, compared with projection onto the PC1 axis in **A**.

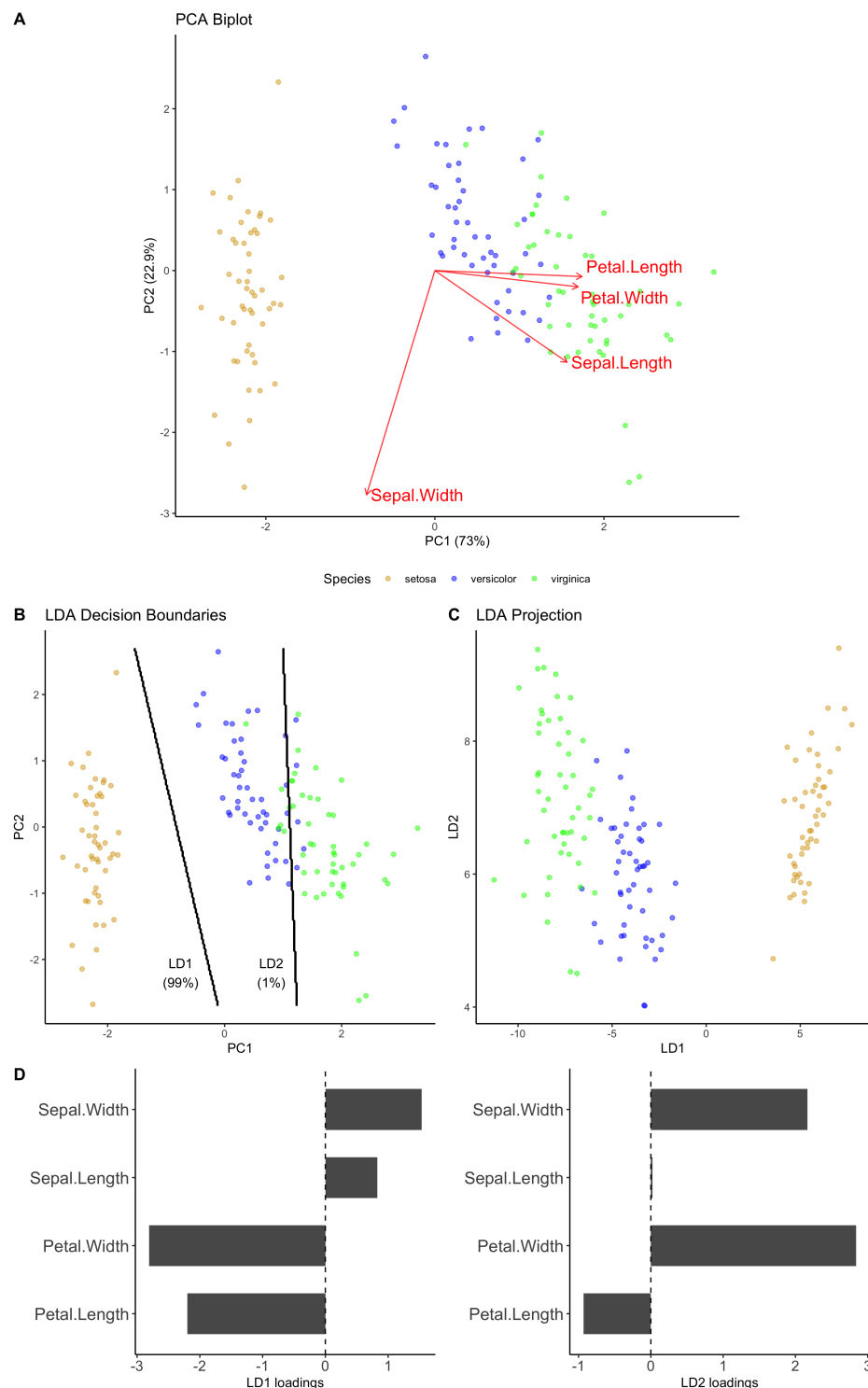


Fig. 2.10 Representation of loadings in principal component (PCA) and linear discriminant (LDA) analysis using Fisher's iris data, which contains the measurements of three different iris flower species. **A** A biplot showing the loadings of how the four variables contribute to the first two principal components. (*Note these loadings have been scaled up for clarity*). **B** shows the LDA-based decision boundaries (discriminant axes) represented in the PC1 and PC2 subspace. The actual boundaries are hyperplanes in the four dimensional space of the original data. **C** shows the data projected onto these LD axes and how they serve to separate the labelled groups. **D** shows the relative loadings of variables with respect to each linear discriminant. PC: principal component. LD: linear discriminant

The advantages of using LDA over other methods of classification are that the decision boundaries that it creates are of low variance and stable compared with the decision boundaries estimated by more complex classification methods (random forests, support vector machines).¹²⁸ Additionally, LDA implements classification problems with multiple groups in a straightforward and easy to interpret with respect to the loadings that it estimates. The goal of this project was to understand the differences between identified sub-groups of patients, not create a model that may predict with high accuracy but might be difficult to interpret.

The implementation of LDA in the *MASS* library in R allows the user to fit an LDA model with multiple labels and calculate the relative importance of the contributing features.¹²⁹ The scaling coefficients can be used to linearly transform the data matrix so that the points can be projected onto a new linear discriminant axis as shown in Figure 2.10c.

2.5.2 Assessment of LDA model performance

Several strategies were considered to assess LDA model performance. The validation test set approach involves splitting the data into a training set and testing set by randomly selecting 70% of samples for training. The model is fitted using the training data and then tested on the unseen (hold-out) testing data. The fitted model is used to predict cluster labels using the testing data and these can be compared with the actual cluster labels to determine the accuracy of the model. The classification error can also be reported using the area under the receiver operating characteristic curve (AUROC).

This model validation approach works well for data with a large number of samples. The accuracy estimates can vary with smaller sample sizes depending on how the training and testing sets are sampled. If borderline cases are under-represented in the training set, then the model may fail to classify borderline cases correctly in unseen testing data. This phenomenon is often referred to as over-fitting.

To obtain accurate estimates of model performance and avoid over-fitting, three approaches are possible. One approach, called bootstrapping, involves resampling the data points with replacement and calculating the average accuracy of the model over hundreds or thousands of iterations. Each iteration uses a different random sample of training and testing data points.

The second approach is called k -fold cross-validation. The data is split into a number (k) of subsets, and the hold-out method is repeated k times. The average performance across all k -folds is reported. This approach is computationally expensive but is feasible on relatively small data.

The third approach, called leave-one-out cross-validation, involves removing a random data point and predicting the class of the removed point using the rest of the data. This is repeated for $k-1$ times, where k is the number of samples in the data set. It amounts to an extreme form of k -fold cross-validation and is best suited for smaller data sets due to its high computational cost.

For the analysis presented in this thesis, leave one out cross-validation was used due to the relatively small cluster sizes. The *caret* library in R automates the above process and reports accuracy but there is no associated confidence interval, nor can AUROC be calculated without hold-out testing data.¹³⁰ To ensure that there was an appreciation of the variance of this accuracy statistic and to calculate AUROC, a bootstrapped approach was also taken. This offered an additional way to check model performance but also facilitated reporting of average model accuracy and AUROC statistics with confidence intervals.

2.6 Endotype characterisation

Once endotypes had been identified, clinical variables were used to characterise each endotype. Comparisons between endotypes were made for biochemical, physiological and haematological variables using analysis of variance (ANOVA) with Tukey's *post hoc* test or Kruskal-Wallis test with Dunn's test depending on the individual variable distributions. Binary categorical features (e.g. positive bacterial culture) were compared using logistic regression or the χ^2 test. Ordinal variables (e.g. SOFA score) were compared using the Kruskal-Wallis test.

Patient outcomes were compared using logistic regression, time to event methods (Kaplan-Meier analysis) and where adjustments were made for other measured variables, Cox proportional hazards were estimated.

Cluster stability was determined using the adjusted Rank index (Section 2.2.3). Cluster transitions over time were observed using Sankey diagrams.

2.6.1 Enrichment of gene lists

Microarray probes were linked to genes in the human genome using the *illuminaHumanv4.db* library which contains a database of annotated probes available on the version of microarrays used in this study.

Each gene module, identified by WGCNA, was submitted to online enrichment tools which calculated the relationships between gene lists and known biological pathways. The statistical

over-representation test was used to identify processes and themes for each module. For a given list of random genes one would expect these genes to be associated with a pathways spread proportionately across the genome. If multiple genes from a submitted list are associated with a single process and this is greater than expected compared with a reference list then this pathway was said to be over-represented.

The binomial statistic is used to calculate the significance level of the represented pathway, with the null hypothesis that there is no difference from the expected number of gene in the reference genome. The p value generated by this binomial comparison is sometimes referred to as the hypergeometric p value. This process involved multiple comparisons to be made simultaneously (one for each potential pathway). Hypergeometric p values were corrected using the Benjamini-Hochberg method.

Each gene module was labelled with the over-represented pathway with the highest significance level attributed to it. Where no known pathway was significantly associated with a gene list, the module was labelled 'no significant pathway'.

There are a number of platforms and tools available for enrichment of gene lists. In this thesis, metaspape [<https://metaspape.org>] and enrichR (<https://maayanlab.cloud/Enrichr/>) were both used as they both simultaneously collate the results from a number of gene ontology databases (KEGG, GO) and pathway databases (reactome).^{131,132}

2.6.2 Determination of differential gene expression

Microarrays were designed to compare the expression of genes between tissue samples that were subject to different experimental conditions prior to extraction of mRNA. A standard workflow for differential gene expression analysis usually involves the following steps:

- Quality control and normalisation steps outlined above in section 2.3.1.
- Linear model fit to experimental model conditions for every sample in a binary, model matrix arrangement called a 'design' matrix. Each column of this matrix is an experimental condition and will form a new term in a linear regression model. Each row is a sample.
- A second 'contrasts' matrix which uses the design matrix to create a new matrix based on which contrasting experimental conditions or disease state are to be compared. For example, comparing endotype 'A' and endotype 'B'. Simple experiments do not require a contrast matrix.

- The linear model is fitted similarly to standard linear regression, using linear algebraic methods.

Comparing gene expression between two conditions using a linear model amounts to comparison of means using a t-test. Moderation of the t -statistic is necessary as microarray data tend to have a low number of replicates and are subject to multiple tests of a large number of genes with limited variance estimates. Shrinkage of variances using the background levels of other genes (V_0) gives better estimates of variance (\hat{V}_g) for a given gene (g) with variance V (equation 2.5).

$$\hat{V}_g = V_0 + V_g \quad (2.5)$$

The *limma* R package uses an empirical Bayes method to moderate the t-statistic, which in addition to shrinking the variances of the residual values from the linear model, provides extra degrees of freedom. The overall affect of these methods is to produce more robust linear models which determine differential gene expression more reliably. These moderation methods have been validated using spike-in control experiments and simulations.^{118,133}

The results from application of the above statistical methods are presented in a table containing estimates of the following values for each probe:

- the \log_2 fold change
- B-statistic, which represents the log-odds of a gene being differentially expressed
- calculated p value
- Benjamini-Hochberg adjusted p value (false discovery rate).

The results from this table were presented in a scatter plot, often referred to as a ‘volcano’ plot, which highlighted genes that had statistically significant differential expression based on their false discovery rate (FDR) and \log_2 fold change. Genes identified as significant were submitted for pathway enrichment as described in section 2.6.1 to determine the biological processes that might be important in different sub-types.

2.7 General statistical methods

All analysis was conducted in the statistical language R version 3.6.2.¹³⁴ The list of packages used include: *dplyr*, *ggplot2*, *WGCNA*, *MASS*, *survival*, *survminer*, *ggfortify*, *cowplot*, *lumi*, *limma*, *illuminaHumanv4.db*, *networkD3*, *htmltools*.^{117,135–143}

For descriptive data results are presented as mean with standard deviation, or median with inter-quartile range. Comparisons were made using Student's t-test for normally distributed data and Wilcoxon-rank sum tests for non-normal data. Where more than two groups were compared, ANOVA with Tukey's *post hoc* test was used for normally distributed variables and Kruskal-Wallis with Dunn's test for non-normally distributed variables. Time to outcome data was analysed using the Kaplan-Meier method. If multiple covariates were used the Cox proportional hazards were estimated. Pearson's correlation coefficient was used to measure correlations between variables. Heatmaps were used to assess the clusters and cells were coloured by the z-score values of each variable. Correlation heatmaps used Pearson's correlation coefficient values to colour the cells. The significance threshold used was $p < 0.05$ and where multiple comparisons were made the false discovery rate (FDR) was used for correction.

Schematics and figures were produced using Lucidchart (www.lucidchart.com) and Adobe Illustrator (Adobe Inc, CA, USA). Result plots were created using R.

CHAPTER 3

Clustering of biological data

3.1 Overview of results

The results section of this thesis is divided into three separate chapters: clustering, integration and characterisation. These represent each stage of the bioinformatic analysis. A detailed overview of these steps was shown in Figure 2.1. A simplified version is shown in Figure 3.1.

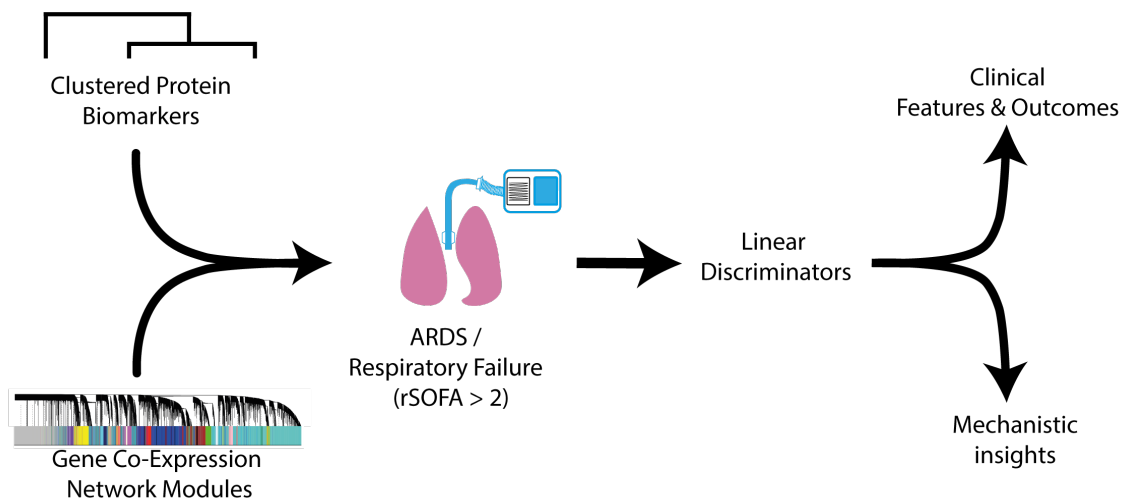


Fig. 3.1 Simplified schematic of the data analysis workflow of this thesis. Protein biomarker values were combined with explained variance values of gene modules calculated by WGCNA. This combination of features was used to discriminate the differences between patients with severe respiratory failure / ARDS in each cluster using an LDA model. Key gene modules were enriched to identify the dominant biological processes and pathways. Clinical information was then integrated in order relate mechanism to clinical features and patient outcomes.

3.2 Hierarchical clustering of protein biomarkers

Soluble immune mediators are proteins released by both non-immune cells in response to damage or recognition of infection, and by cells directly involved in the immune response to infection and injury. An ever-increasing number of immune mediators are recognised for the role in recruitment and modulation of immune cells. Collectively these proteins are called cytokines. Cytokines are generally grouped based on the types of immune cell that release them or they recruit. For example, type 1 interferons (IFN- α and IFN- β) are associated with an anti-viral immune response, whilst TNF- α is associated with granulocyte-mediated immune responses. Not all immune mediators that influence immune function are cytokines: granulocyte-macrophage colony-stimulating factor (GM-CSF) is a growth factor affecting granulopoiesis that also functions as a cytokine.

The aims of this step of the analysis was to determine if there were different groupings of patients (clusters) with similar concentrations of protein biomarkers. Similarities between independently-derived clusters might infer stereotyped immune responses in the context of critical illness.

3.2.1 Preprocessing and imputation

The data provided for this study consisted of measurements of 26 protein biomarkers in 199 samples for the GAINs study and 34 protein biomarkers in 378 samples for the MOSAIC study. For the MOSAIC study, there were technical replicates of IL-8 assays. The mean value across both replicates was used. The protein biomarker data in the MOSAIC study included 68 samples taken during convalescence. 310 samples remained after exclusion. Further exclusions for samples with rSOFA score less than two or with more than three missing values, meant 154 samples from the MOSAIC study remained for cluster analysis.

Measured protein biomarker concentrations were log-transformed and scaled (centred on the mean) prior to analysis as the distribution of these data were right skewed. Scaling improves the performance of Euclidean distance-based clustering and allows for integration with other, similarly scaled variables in downstream analysis. The proportion of missing data was equal to 0.6% for the GAINs samples and equal to 0.9% for the MOSAIC samples. Where data was missing in a given sample, generally, only one or two values were absent. It was considered low risk to impute these missing values instead of discarding all the data from these samples entirely.

The adjusted Rand index was used to assess the performance of different imputation methods. Table 3.1 shows that imputation with the mean performed better than MICE-based imputation.

		Complete data clusters		
		1	2	3
Mean imputed data clusters	1	96	0	0
	2	0	59	0
	3	0	0	21
Adjusted Rand index = 1				
		Complete data clusters		
		1	2	3
MICE imputed data clusters	1	92	2	2
	2	1	58	0
	3	0	0	21
Adjusted Rand index = 0.91				

Table 3.1 The effect of different imputation strategies (mean, MICE) on cluster assignment using hierarchical clustering for the protein biomarker Values from the GAINs study. ‘Complete data clusters’ refers to clustering of samples with no missing values ($n = 176$). The hierarchical clustering dendrogram was cut to generate three clusters. The same clustering methods were used on the protein biomarker value with missing data ($n = 199$), after imputation. The imputed data cluster assignments were compared with complete data clusters assignments for the in-common samples ($n = 176$). Concordance was evaluated using the adjusted Rand index. Imputation using the mean had perfect concordance (adjusted Rand index = 1) compared with imputation using MICE (adjusted Rand index = 0.91).

3.2.2 Hierarchical clustering: linkage methods

Four different linkage methods were available for hierarchical clustering: Ward's, average, single and complete. Projection of the protein biomarker values into the principal component (PC) space demonstrated no obvious separation of data points into distinct clusters. Hierarchical clustering methods served to segment these data points into different groups. The effect of each linkage method was assessed visually using pair-wise scatter plots of the first three principal components. The data points were segmented into increasing cluster divisions (k) to see the effect of data segmentation as k increased. Figures 3.2 and 3.3 demonstrate how each of the linkage methods segmented the protein biomarker values into three or four groups. On this basis, the Ward linkage method of hierarchical clustering was chosen as it produced consistent segmentation of data points that could be easily interpreted and visualised.

Inclusion of all patients in clustering

The method of hierarchical clustering assigns every data point to a cluster. This may result in outlier samples becoming incorporated into these clusters which might influence downstream analysis. Alternatively, outliers may be assigned to a separate cluster by the linkage method, in which case clustering provides a useful method for excluding outlier samples. It can be seen from the principal component plots (Figures 3.2 and 3.3) that no samples required exclusion, nor was it the case that clustering identified an outlier group as a distinct cluster in these data.

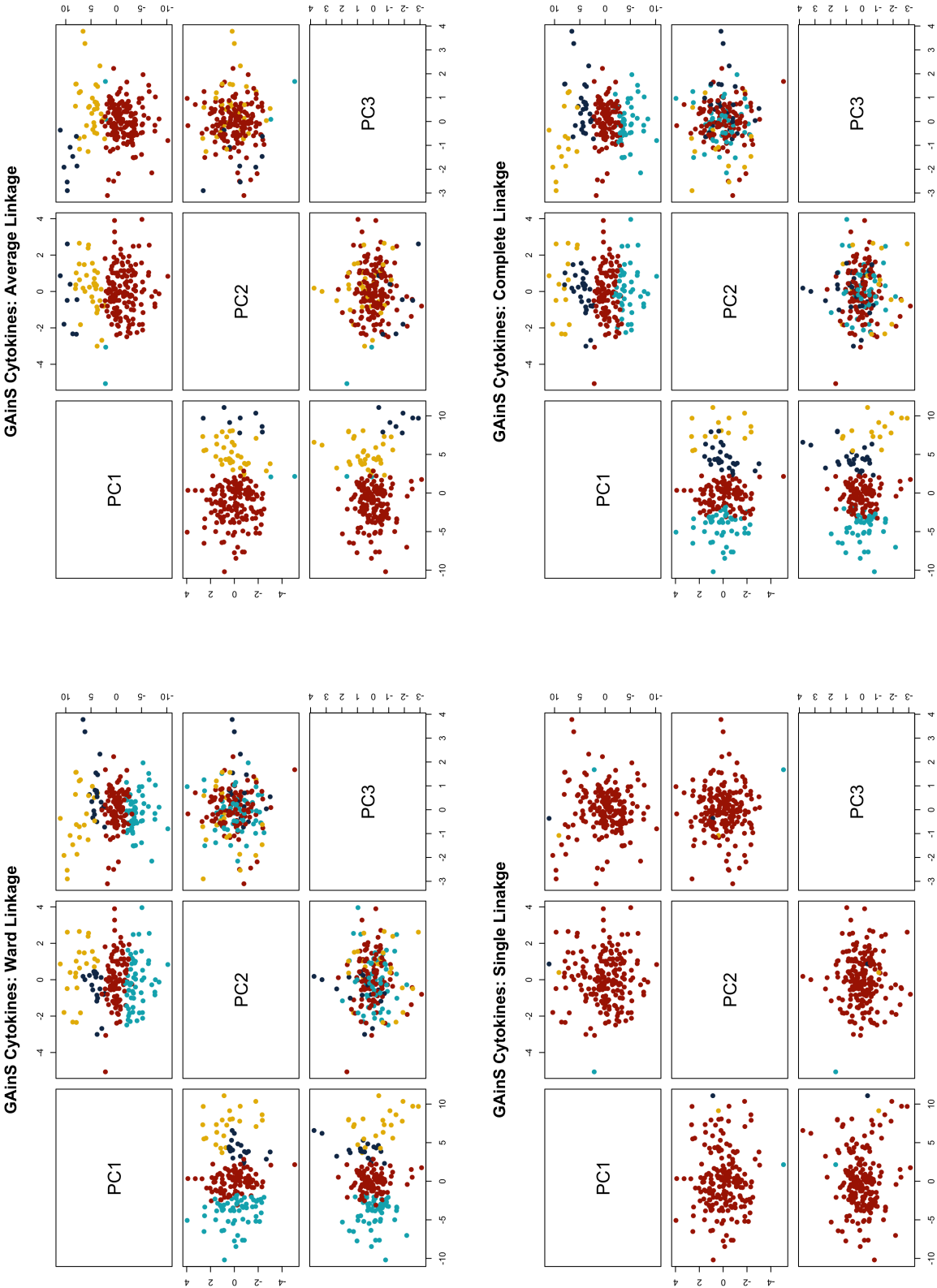


Fig. 3.2 (Caption on page following Fig 3.3.)

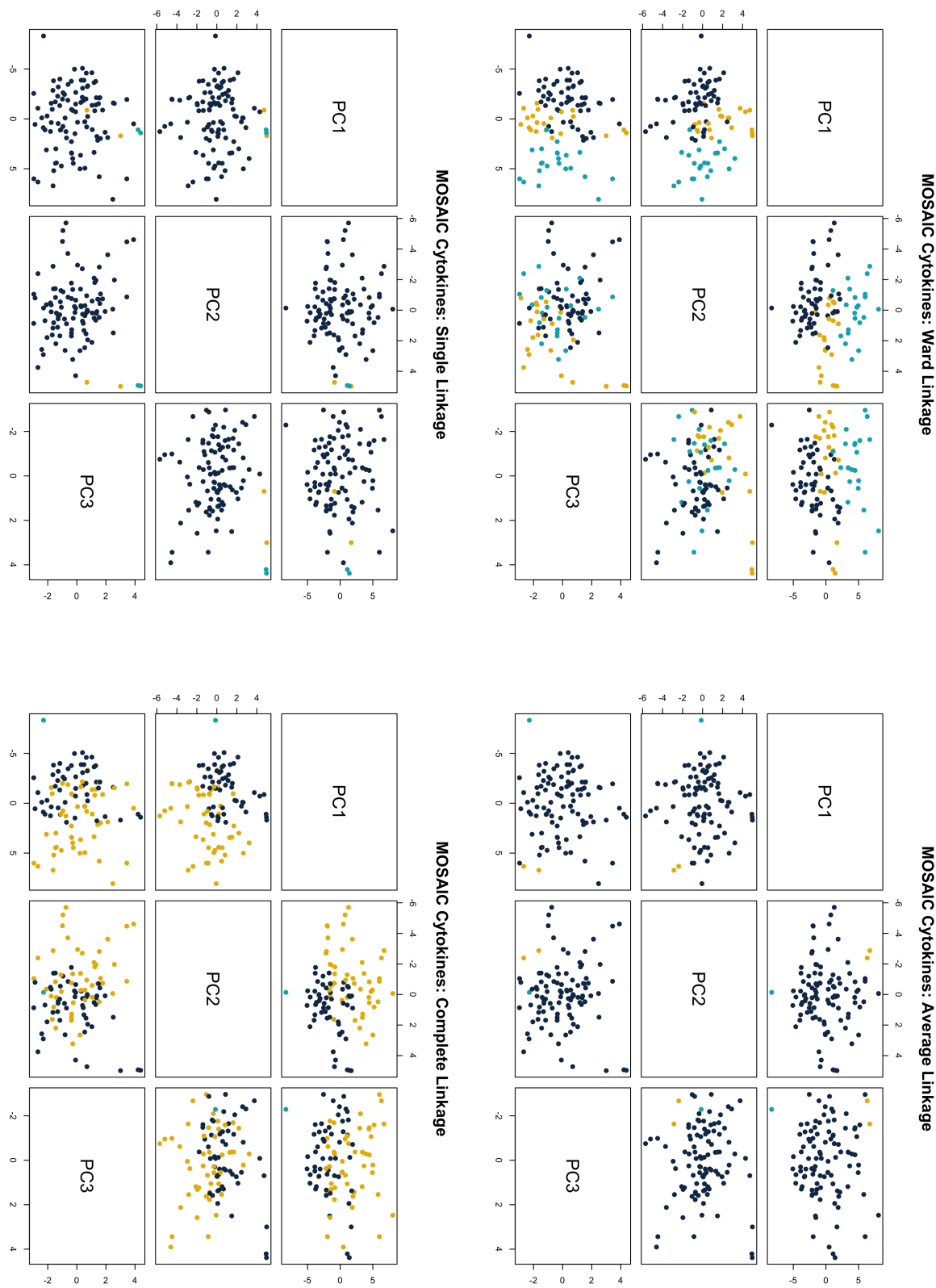


Fig. 3.3 (Caption on following page.)

Figures 3.2 and 3.3 Visualisation of the segmentation of protein biomarker data in the principal component space using different hierarchical clustering linkage methods. Each point is a study sample which represents the combined results from 26 (GAinS) or 35 (MOSAIC) protein biomarker assays, projected into the principal component space. Only the first three principal components (PC) are shown here. Colours indicate cluster assignment. Three or four clusters were chosen to emphasise the effect of linkage methods on clustering. The Ward and complete linkage methods produced similar cluster assignments with the GAinS protein biomarkers, but this was not the case with the MOSAIC protein biomarkers. The single and average linkage methods did not produce easy to interpret cluster assignments for either set of protein biomarker results. Ward linkage was chosen as the preferred linkage method for hierarchical clustering.

3.2.3 Hierarchical clustering of protein biomarker profiles from patients recruited to the GAINs study identified three clusters

Hierarchical clustering using Euclidean distance and Ward linkage gave an optimum cluster number of three for protein biomarker profiles from patients in the GAINs study. This number was derived from a combination of inspection of the dendrogram, k-means elbow method, the *NbClust* cluster metrics (Figure 3.4) and heatmap visualisation (Figure 3.5). The number of samples and mean protein biomarker z -scores in each cluster are listed in Table 3.2.

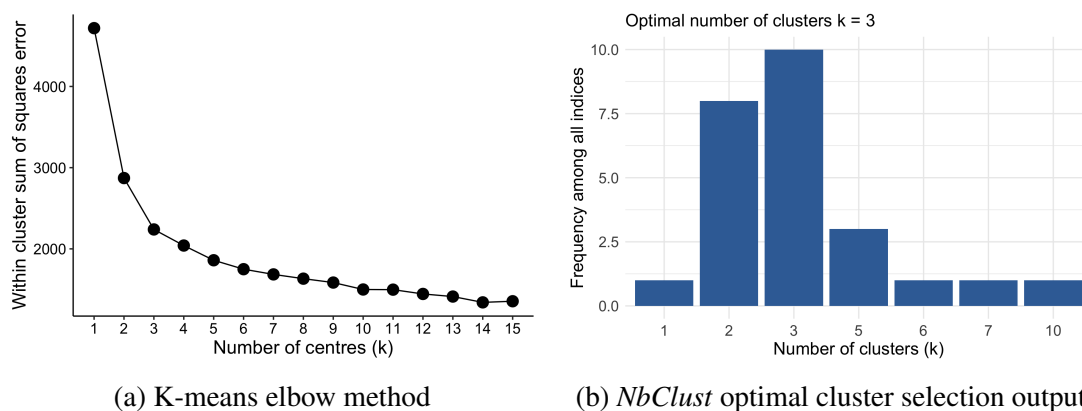


Fig. 3.4 Optimal cluster determination for the protein biomarker values in the GAINs study using (a) k-means elbow method and (b) the majority voting methods from *NbClust*. Both methods suggested that three clusters was the optimal number for these data.

The mean, log-transformed values of each measured protein biomarker for each cluster group are shown in Figure 3.6. The line plot and the heatmap (Figure 3.5) both demonstrate the polarised distributions of protein biomarkers observed in these three clusters. Each cluster had statistically significant different concentrations for all measured protein biomarkers (Appendix D.1). The concentrations of the protein biomarkers in the ‘purple’ cluster were on average, universally elevated whilst the concentrations in the ‘green’ cluster were, on average, universally depressed. The ‘yellow’ cluster had relatively uniform concentrations that approximated the means of the log-scaled sample values. The polarised nature of the identified clusters suggests there is severe immune dysregulation in patients with sepsis, who may have globally elevated or depressed cytokine and chemokine concentrations.

Figure 3.5 also shows how patients with a diagnosis of ARDS are distributed across all three clusters. This demonstrates the heterogeneity, at an immunological level, of ARDS in the context of sepsis.

Protein biomarker	Yellow cluster (1)	Purple cluster (2)	Green cluster (3)
n	98	59	42
CCL3 (MIP-1 α)	0.1014	0.7788	-1.3308
IL-1 β	0.0174	0.9148	-1.3256
IL-2	-0.0826	0.8233	-0.9638
IL-4	-0.0554	0.7844	-0.9726
IL-5	0.0630	0.7961	-1.2653
CXCL10 (IP-10)	-0.1105	0.8403	-0.9225
IL-6	0.1348	0.7530	-1.3722
IL-8	0.0189	0.8824	-1.2836
IL-10	0.0520	0.8191	-1.2720
CCL11 (Eotaxin-1)	-0.0147	0.8259	-1.1260
IL-12p70	-0.0691	0.8523	-1.0360
IL-13	0.0371	0.7734	-1.1732
IL-17A	-0.0175	0.7364	-0.9936
IFN- γ	0.0235	0.9216	-1.3495
GM-CSF	0.0834	0.8496	-1.3880
TNF- α	0.0903	0.8988	-1.4733
CCL4 (MIP-1 β)	0.0058	0.8707	-1.2367
IFN- α	-0.0539	0.7544	-0.9340
CCL26 (Eotaxin-3)	0.0900	0.8396	-1.3895
CCL2 (MCP-1)	0.0477	0.5878	-0.9370
CXCL11 (I-TAC)	0.0508	0.8669	-1.3364
CXCL9 (MIG)	-0.0592	0.9853	-1.2459
TNFR-2	-0.0510	0.8936	-1.1362
CCL22 (MDC)	-0.1504	0.8793	-0.8842

Table 3.2 Mean z scores of protein biomarkers from samples in each GAINs cluster. Samples from the ‘purple’ cluster had globally raised protein biomarker concentrations whilst samples from the ‘green’ cluster had globally suppressed protein biomarker concentrations. All three clusters appear to have, on average, polarised responses. The relative concentrations and distributions of each measured protein biomarker in patients from each cluster are shown in Appendix Figure D.1.

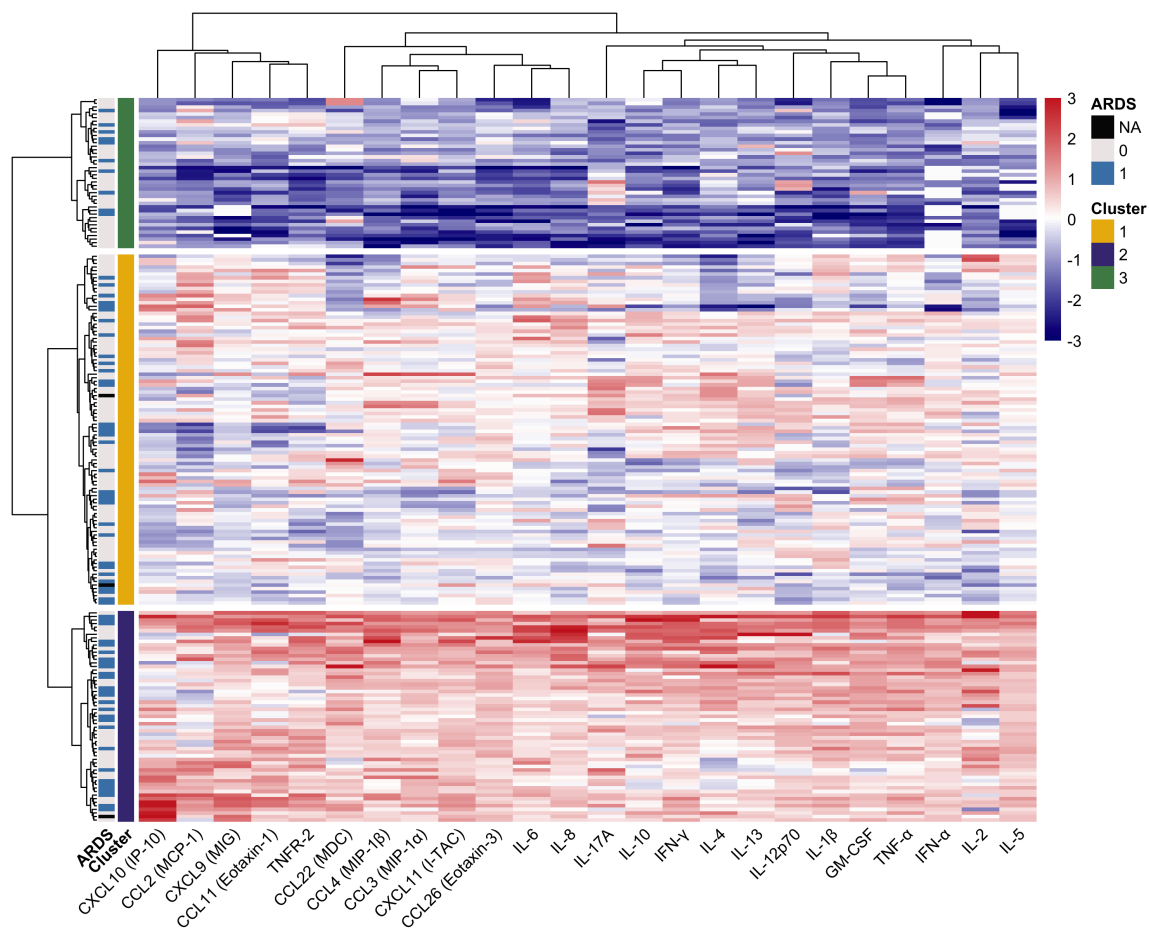


Fig. 3.5 Heatmap showing the relative values of measured protein biomarkers, on the z scale (zero-centred), for GAINs samples in each cluster. The cluster dendrogram derived by Ward linkage method is on the left side of the heatmap, dividing the samples into the clusters labelled by their colour bars, or ARDS status. It can be observed that there were patients with ARDS in each of the three clusters (blue and grey bands). NA: missing value

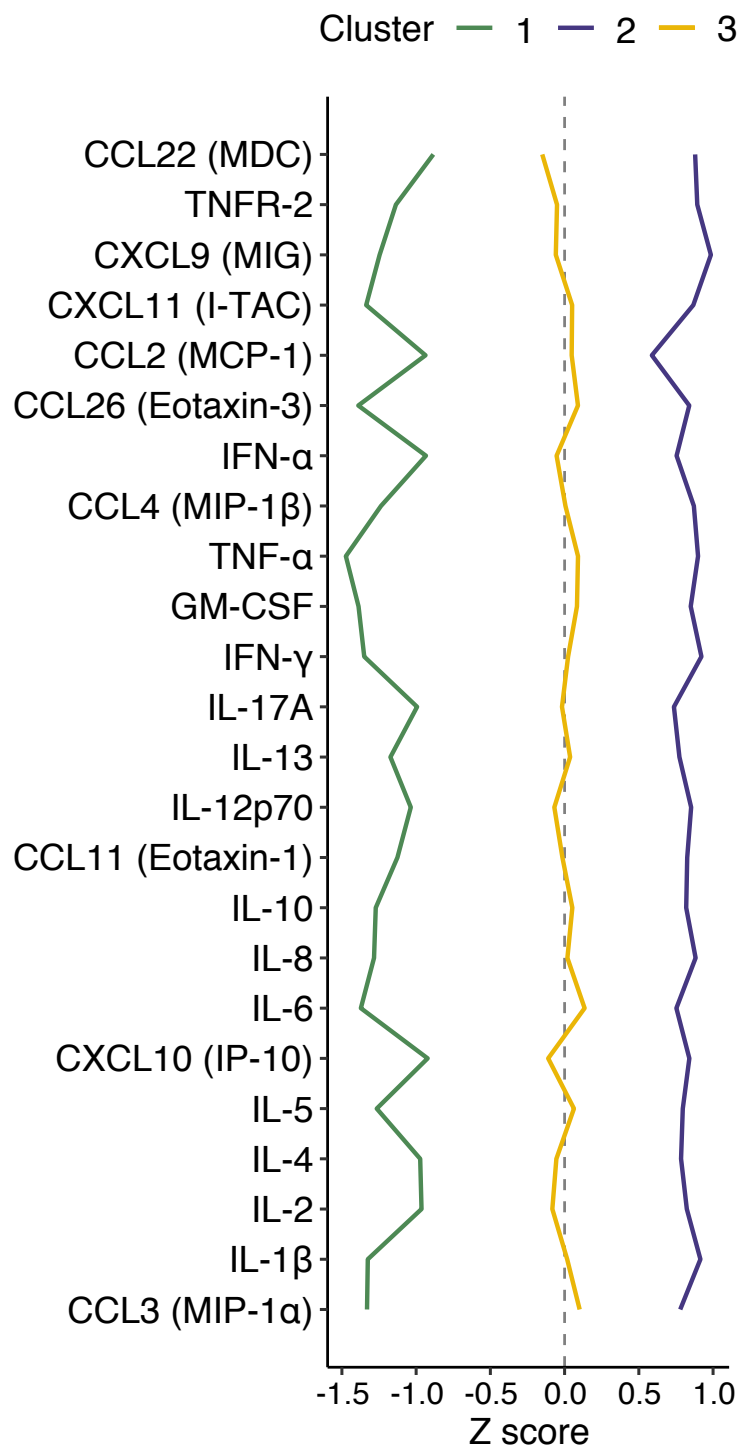


Fig. 3.6 Mean protein biomarker values in samples from each GAINs cluster. The three clusters were characterised by polarized distributions of protein biomarker concentrations. This is another representation of how different each of the clusters were. Cluster 2 (purple) samples had significantly elevated concentrations of all measured cytokines and chemokines, suggesting a dysregulated, excessive cytokine response involving both pro- and anti-inflammatory cytokines. Cluster 3 (green) had significantly lower mean concentrations of all measured protein biomarker suggesting pan-suppression of cytokine responses.

3.2.4 Hierarchical clustering of protein biomarker profiles from patients recruited to the MOSAIC study identified three clusters

Hierarchical clustering using Euclidean distance and Ward linkage gave an optimum cluster number of three for the protein biomarker profiles from the MOSAIC study. This number was derived from a combination of inspection of the dendrogram, k-means elbow method, *NbClust* cluster metrics, (Figure 3.7) and heatmap visualisation (Figure 3.8). The number of samples assigned to each cluster are listed in Table 3.3.

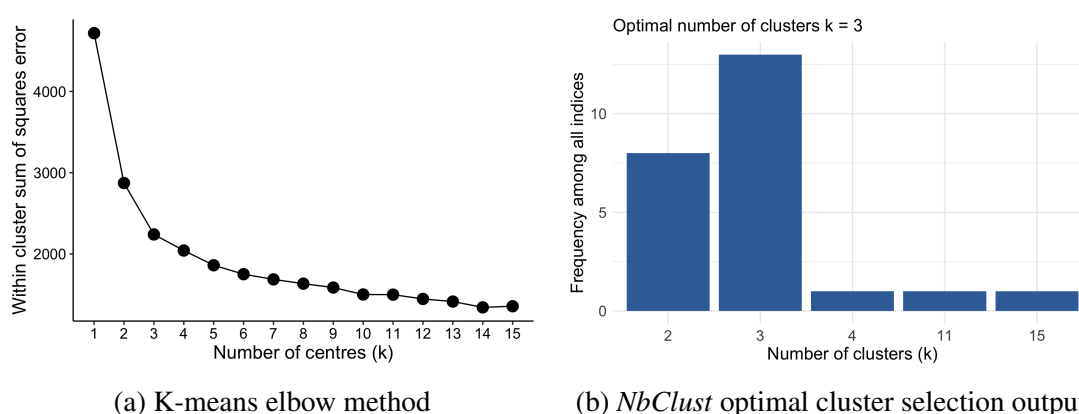


Fig. 3.7 Optimal cluster determination for the immune mediator values from the MOSAIC study using (a) k-means elbow method and (b) the majority voting methods from *NbClust*. Both methods suggested that three was the optimal number of clusters.

Samples from the 'red' cluster (3) were associated with much higher concentrations of inflammatory mediators associated with the acute phase response (IL-6, TNF- α). 'Red' cluster samples also had significantly higher concentrations of IL-15. IL-15 is associated with severe disease in influenza which is thought to be mediated by NK cell activation.¹⁴⁴ Samples from the 'grey' cluster (2), which represented the largest group, were associated with lower concentrations of these three mediators but the highest levels of of interferon- α 2a (IFN- α 2a). IFN- α 2a is a type I interferon and is associated with an anti-viral immune response. Samples from the 'blue' cluster (1) had moderately raised levels of TNF- α . IL-15 and IL-6 and depressed levels of IFN- α 2a similar to samples from the 'red' cluster (3). The immune mediator profiles in the patients from the MOSAIC study were generally less polarised, in comparison with the clusters identified in the GAINs study.

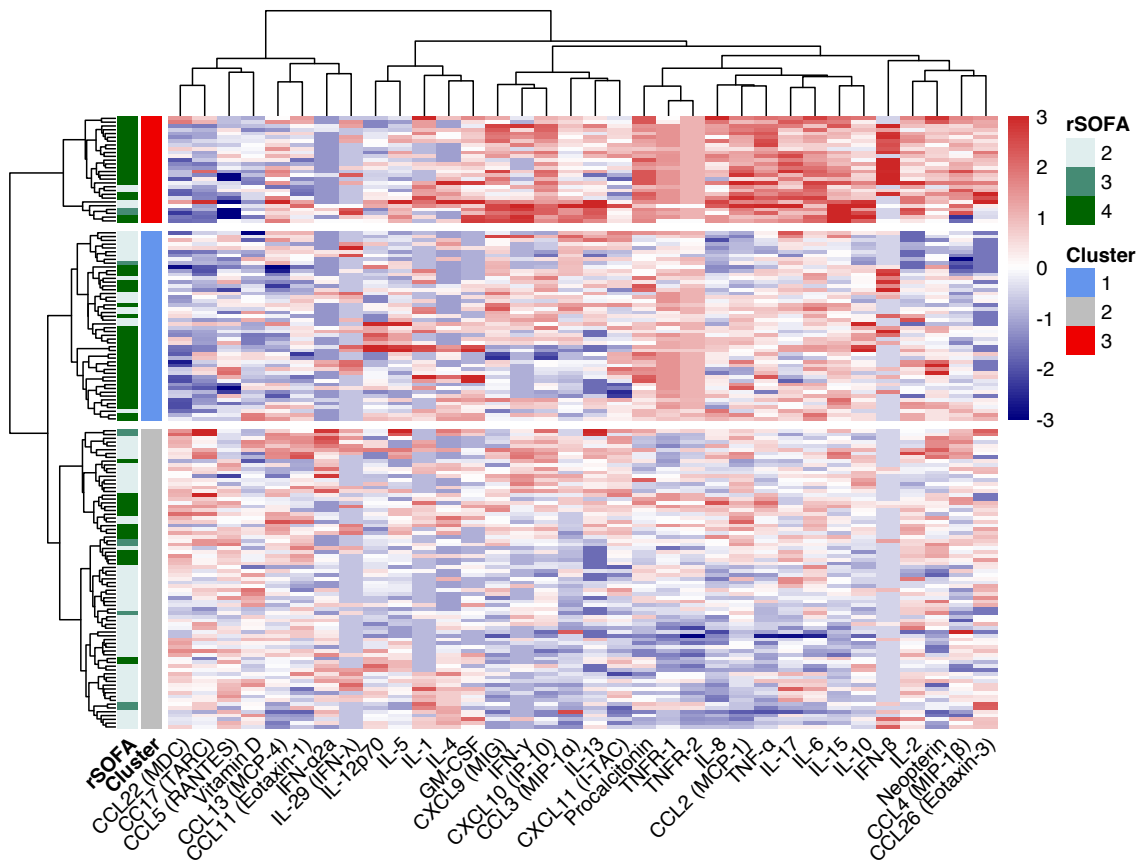


Fig. 3.8 Heatmap showing the relative values of measured immune mediators, on the z scale, for MOSAIC samples in each cluster. The cluster dendrogram derived by Ward linkage method is on the left side of the heatmap, dividing the samples into the cluster labelled by their colour bars. The colour bar labelled 'rSOFA' shows the respiratory SOFA scores for each patient. The dark green boxes indicate samples from patients on mechanical ventilation ($\text{rSOFA} \geq 3$). These patients are distributed across all three clusters, although a higher proportion of them are found in the 'red' and 'blue' clusters.

Immune mediator	Blue cluster (1)	Grey cluster (2)	Red cluster (3)
n	50	79	28
CCL5 (RANTES)	-0.2236317	0.29514907	-0.9229394
IFN- α 2a	-0.5764506	0.26389183	-0.8822796
CCL22 (MDC)	-0.882019	0.42218442	-0.7077652
CC17 (TARC)	-0.6928466	0.37227819	-0.6150973
Vitamin D	-0.366662	0.05966778	-0.5413645
IL-12p70	0.14141123	-0.1314306	-0.227186
CCL11 (Eotaxin-1)	-0.1549202	0.15225921	-0.1057977
CXCL11 (I-TAC)	0.12904854	0.00264832	0.1483144
IL-29 (IFN- λ)	0.11410136	-0.0596924	0.1551903
IL-5	0.13120274	-0.027955	0.286988
IL-4	-0.0851889	0.09144718	0.3159596
CCL13 (MCP-4)	-0.3247548	0.13225012	0.6277915
IL-1	0.25574758	-0.0122592	0.6641887
CCL3 (MIP-1 α)	0.13404839	-0.3541256	0.7035267
Neopterin	0.33853049	0.03459873	0.8441284
IL-13	-0.0969028	-0.1374579	0.8840543
IFN- γ	-0.0778043	-0.1140423	0.8858542
GM-CSF	0.13160948	-0.0577402	0.9551785
CCL4 (MIP-1 β)	-0.4020597	0.07512544	0.9799912
TNFR-2	0.83911253	-0.3440906	1.0067219
CCL26 (Eotaxin-3)	-0.6345743	0.29252696	1.019404
IL-2	-0.1310917	0.2854464	1.0991339
IL-10	0.25640962	-0.0990634	1.1086215
IL-8	-0.1551415	-0.3011357	1.1266713
TNFR-1	0.89429784	-0.32281	1.2625966
CXCL9 (MIG)	0.32551139	-0.2373243	1.3765793
IFN- β	0.3699887	-0.2387985	1.5182616
CXCL10 (IP-10)	0.04823145	-0.0203755	1.5360618
IL-17	0.24800508	-0.074908	1.5837885
CCL2 (MCP-1)	0.09823013	0.05258789	1.6686114
Procalcitonin	0.6891961	-0.1782709	1.679158
IL-6	0.54364565	0.0336591	1.690874
IL-15	0.51291068	-0.2191639	1.7250176
TNF- α	0.27617064	-0.2659672	1.8052933

Table 3.3 Mean z scores of immune mediator concentrations in patients from each MOSAIC cluster. Samples from the ‘red’ cluster were associated with significantly higher concentrations of TNF α , IL-6, IL-15 and procalcitonin than the other two clusters. Samples from the ‘grey’ cluster were associated with higher concentrations of CCL5, IFN- α 2a and CCL22 than the other clusters. The concentration and distributions of immune mediators in patients from each cluster is also shown in Appendix Figure D.2.

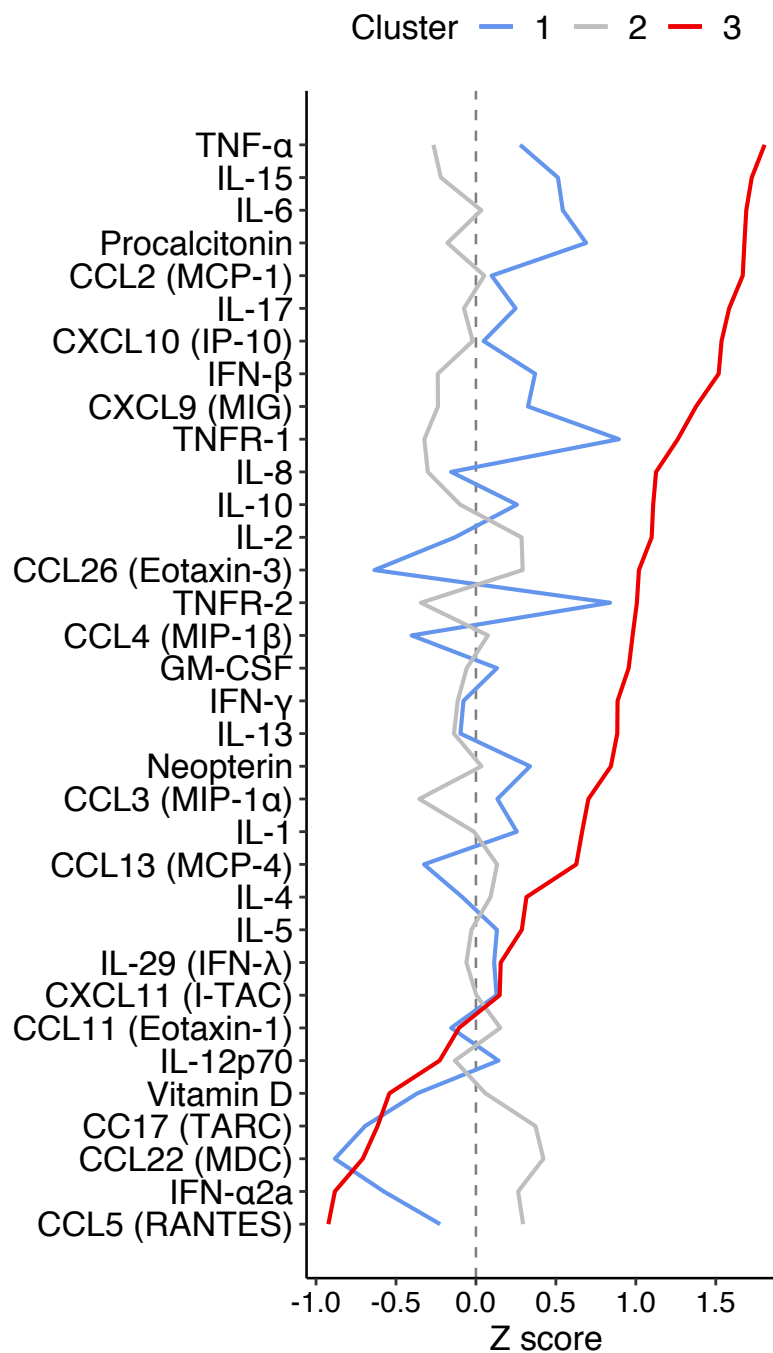


Fig. 3.9 Mean immune mediator value in samples from each MOSAIC cluster. Samples from the the 'red' cluster (3) had significantly elevated concentrations of cytokines and chemokines associated with granulocyte activation. The other two clusters had a mixed, more moderated response that was closer to the sample mean for the majority of measured immune mediators.

3.2.5 Hierarchical clustering of biomarker profiles in the HARP-2 patients did not identify an optimal number of clusters

The HARP-2 researchers measured six protein biomarkers in 511 patients on the day of randomisation. Biomarker samples were taken in some patients at days 3, 7, 14 and 28 post randomisation but these were inconsistently collected and measured. In addition to the cytokines IL-6 and sTNFR-1, the study team measured concentrations of four protein biomarkers: angiopoietin-2 (Ang-2), matrix metalloproteinase-8 (MMP-8), soluble receptor for advanced glycation end products (sRAGE), surfactant protein-D (SP-D). These protein biomarkers have previously been shown to be associated with worse outcomes in patients with ARDS (Section 1.5.1).

Of the six mediators measured by the HARP-2 study team, only IL-6 was in common with the other two studies in this project.

There was inconsistent coverage of biomarker measurements at times other than day 1 (recruitment). If sampling days 1, 3 and 5 were considered, only four biomarkers had been measured on these three days, and only in a much smaller sample of patients. Stable clusters across these three sampling times could not be demonstrated due to the paucity of features. For these reasons only recruitment day samples were considered for clustering of biomarker profiles.

The k-means elbow method identified no clear candidate value of k consistent with the start of a plateau for the within cluster sum of squares values (Figure 3.10a). This suggested that there was no optimum cluster that could be identified using k-means. The *NbClust* package determined that either two or six as optimum cluster numbers (Figure 3.10b).

Inspection of cluster dendrograms and heatmaps produced by the clustering algorithms suggested that three clusters was a reasonable compromise and served to recognise distinct properties of these data (Figure 3.12). Both the GAinS and MOSAIC studies had determined three clusters as the optimum number segmentation of protein biomarker results based on compelling evidence. This supported with decision to proceed on the basis of three biomarker-based clusters in the HARP-2 study. To ensure that there was no untoward reason for failure to identify clusters principal component projections of the cytokine concentrations were plotted and reviewed (Figure 3.11).

The number of samples (patients) and relative levels of biomarkers in each cluster are shown in Table 3.4. Samples from the dark green cluster (3) were associated with globally depressed measured mediators (Figure 3.12). Enforcing a three cluster split caused the second large

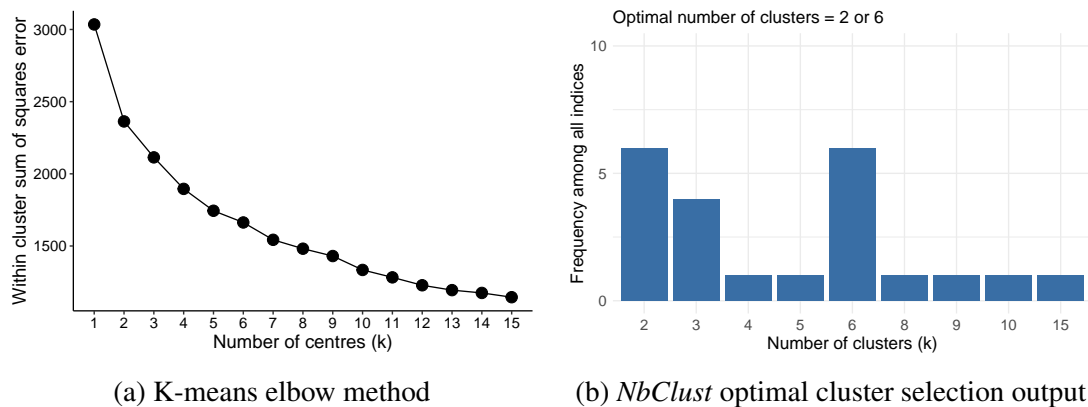


Fig. 3.10 Optimal cluster determination for the protein biomarker concentrations from the HARP-2 study using (a) k-means elbow method and (b) the majority voting methods from *NbClust*. Both methods were inconclusive for an optimum cluster number.

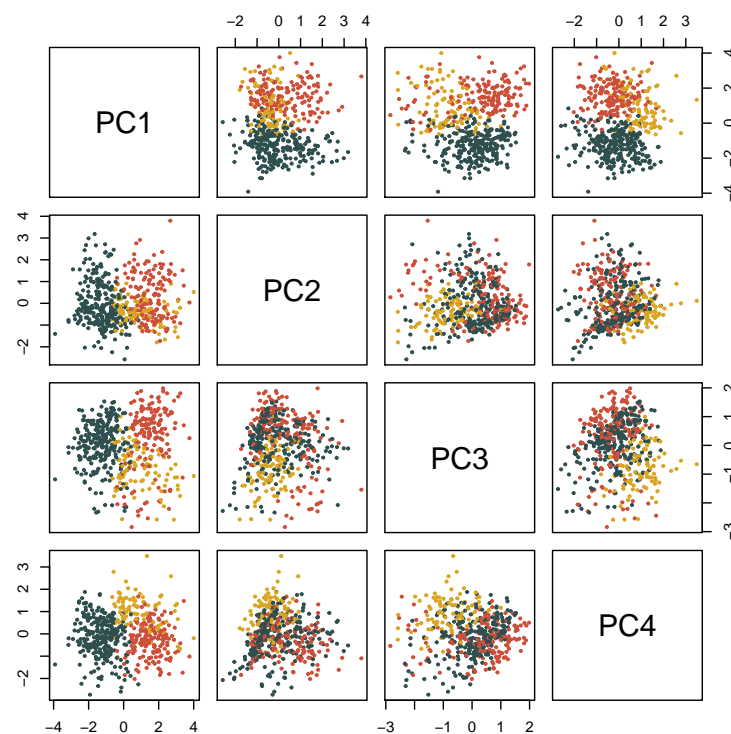


Fig. 3.11 Principal component projection of HARP-2 protein biomarker values. Each point is a patient sample taken at recruitment to the study. Each colour corresponds to a different cluster assignment as determined by the Ward linkage method. Segmentation of the data is consistently observed in the four visualised principal component projections with this method.

branch to divide into two smaller ones ('dark red' and 'dark yellow'). Samples from the 'dark red' cluster (1) were associated with high MMP-8 concentrations. Samples from the 'dark yellow' cluster (2) were associated with high sRAGE concentrations. Both the 'dark red' and 'dark yellow' clusters had similar levels of IL-6 and Ang-2. This suggested that there were two distinct profiles of acute inflammation in patients enrolled in the HARP-2 study, both with evidence of endothelial injury. This distinction might not have been recognised if the samples from these clusters had been merged into a single 'inflamed' cluster, using a two-cluster division of these protein biomarker values.

	Cluster 1 (dark red)	Cluster 2 (dark yellow)	Cluster 3 (dark green)
n	160	89	262
SP-D	0.307	-0.362	-0.064
sTNFR-1	0.56	0.323	-0.451
sRAGE	-0.061	1.438	-0.451
MMP-8	0.663	0.015	-0.41
IL-6	0.652	0.453	-0.552
Ang-2	0.408	0.408	-0.388

Table 3.4 Mean z scores of protein biomarkers in each HARP-2 cluster. Samples from the 'dark red' cluster (1) were characterised by higher MMP-8 concentrations. Samples from 'dark yellow' cluster (2) were characterised by high sRAGE concentrations. Samples from both clusters had high concentrations of IL-6 and Ang-2. Samples from the 'dark green' cluster (3) had depressed levels of all measured biomarkers. The non-transformed concentrations of protein biomarkers in each HARP-2 cluster can be seen in Appendix Figure D.3.

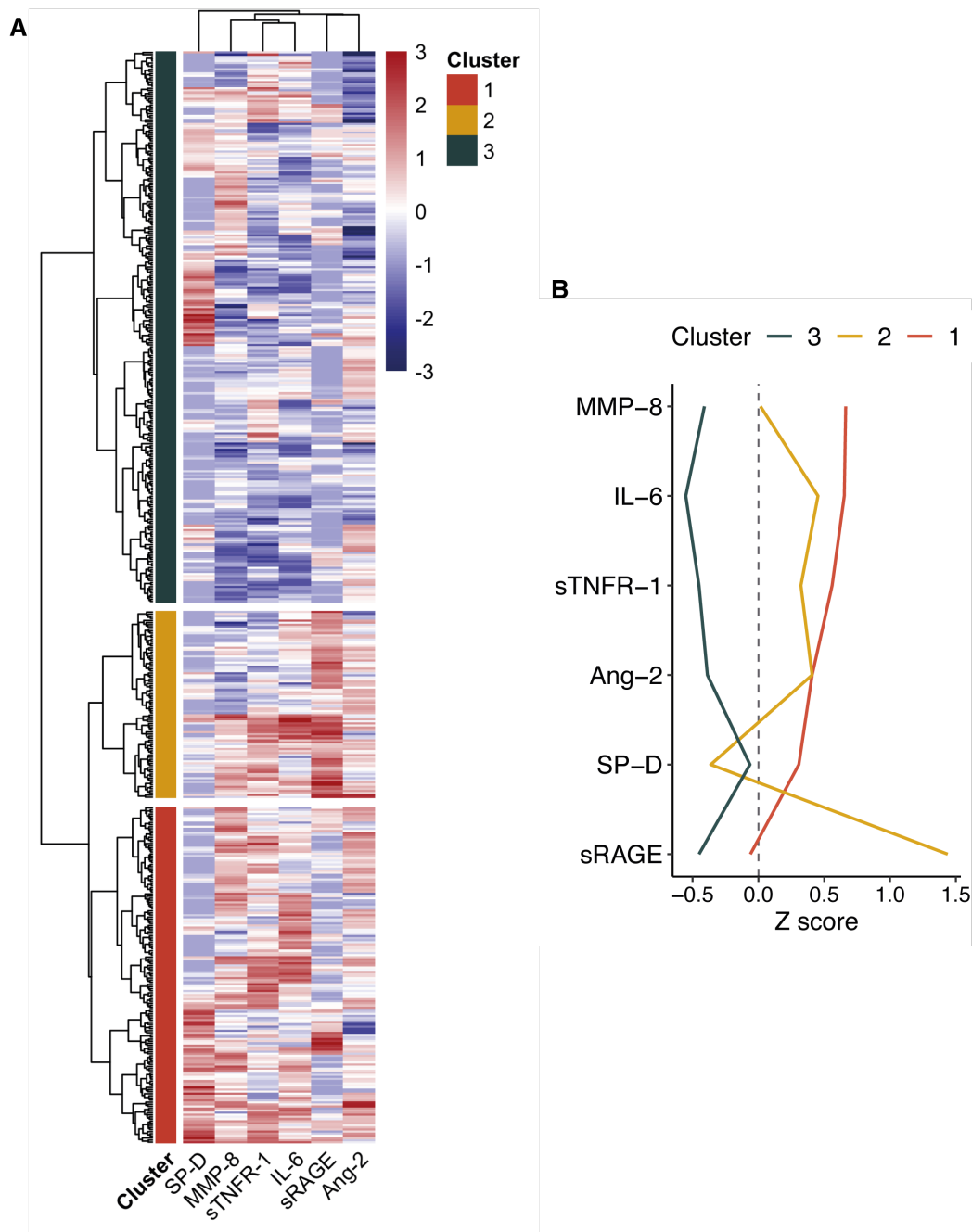


Fig. 3.12 **A** Heatmap showing the relative values of measured protein biomarkers, on the z scale, for the samples in each HARP-2 cluster. The cluster dendrogram derived by Ward linkage method is on the left side of the heatmap, dividing the samples into the cluster labelled by their colour bars.

B Mean protein biomarker values in samples belonging to each cluster from the HARP-2 study. Samples from the 'dark green' cluster (3) were associated with depressed levels of most mediators. Samples from the 'dark red' cluster (1) had higher concentrations of MMP-8. Samples from the 'dark yellow' cluster (2) had the higher concentrations of sRAGE.

3.2.6 Assessment of cluster stability

The adjusted Rand index was used to determine the cluster assignment on asymmetrical partitions of the data which were re-sampled 500 times. The calculated mean, adjusted Rand index values for each study, are listed in Table 3.5. The adjusted Rand index values are consistent with relatively stable clusters in the GAINs and MOSAC studies (ARI > 0.7). Cluster stability was relatively lower for clusters identified in the HARP-2 study (ARI = 0.68).

Study	Samples	Features	Clusters	Adjusted Rand index	95% confidence interval
GAinS	199	24	3	0.83	0.82-0.84
MOSAIC	157	32	3	0.74	0.73-0.75
HARP-2	511	6	3	0.68	0.67-0.69

Table 3.5 Adjusted Rand index values for cluster stability using partitioned, re-sampled data. 'Features' refers to the number of measured biomarkers

3.3 Discussion of protein biomarker clustering

There is a strong argument for the existence of at least three distinct immune profiles, determined by hierarchical clustering of protein biomarkers, in both the MOSAIC and GAINs studies. The 'purple' GAINs and MOSAIC 'red' clusters both demonstrated evidence of dysregulated immunity with raised cytokines in groups associated with innate and adaptive immunity. The polarisation of the relative cytokine concentrations in different patient clusters from the GAINs study was particularly striking.

There are no similarities between the other identified clusters in the MOSAIC and GAINs studies. Samples from the 'green' GAINs cluster had globally depressed cytokines, suggesting immune exhaustion or dysregulated suppression. The 'blue' cluster from the MOSAIC study showed an intermediate (relative to the other clusters) acute phase response with IL-6, TNF- α higher than the mean. Samples from the 'red' and 'blue' MOSAIC clusters had low concentrations of CCL5 and INF- α 2a, which suggested that lymphocyte mediated anti-viral immune responses were less predominant in these patients. Samples from the 'grey' MOSAIC cluster had low concentrations of IL-6, TNF- α and IL-8, consistent with depressed innate immunity. However, this was not a globally suppressed cytokine response, as these patients had higher concentrations of lymphocyte-associated mediators (CCL5, CCL22,

CC17) and higher levels of IFN- α 2a. This suggested a different immune profile in these patients to influenza infection, possibly mediated by lymphocytes and interferons.

The clustering approach to the measured biomarkers in the HARP-2 study was less robust, given that an optimum cluster number could not be determined using the methods to hand. This was likely to be due to the small number of measured biomarkers available. A cluster number of three was arbitrarily imposed on this data based on: inspection of the dendrogram, segmentation of data in the principal component space and because of the number of clusters identified in the GAINs and MOSAIC cytokine data. The calculated adjusted Rand index of a three cluster split, after bootstrapped resampling, was equal to 0.68 (95% CI 0.67-0.69). This value was adequate to demonstrate stability, even though it was lower than the adjusted Rand-index values calculated for the clusters identified in the GAINs and MOSAIC studies.

Two of the three clusters identified in the HARP-2 data were associated with markers of acute innate immunity but with different profiles: 'dark red' was characterised by increased concentrations of IL-6 and MMP-8, 'dark yellow' was characterised by increased concentrations of IL-6 and sRAGE.

It should be noted that in the heatmap figures demonstrating z scores of protein biomarker values for each sample (Figures 3.5, 3.8 and 3.12), only patient samples-based clustering has been taken into consideration. The individual protein biomarkers are also arranged into clusters in these heatmap figures using the same Ward clustering linkage method. The dendrograms showing the relationships between protein biomarkers are visible at the top of each heatmap.

The literature on cytokine responses in influenza or any immune-mediated response often describe 'modules' of cytokines acting in concert.¹⁴⁵ These modules are often assigned themes that relate to immune cells that are influenced by or secrete these cytokines. Performing cluster analysis on cytokine levels alone, not on samples, is of interest but fails to acknowledge the heterogeneity of responses in patients. A cytokine-module approach tends to eliminate the prospect of identifying heterogeneity of immune responses amongst patients. This is because deriving these relationships involves:

- calculation of correlation coefficients for a given cytokine across all samples
- clustering on the values of cytokines that are consistently highly correlated with each other.

Groups of patients that do not fit the themes of the more dominant cytokine relationships will therefore be discarded by this method of analysis. If the above method is applied to the protein biomarker results in the GAINs study the heatmap seen in Figure 3.13 is produced. It can be seen here that there is a cluster of eleven protein biomarkers (IL-1 β , IL-10, TNF- α , CCL26, IL-8, IFN- γ , IL-6, CCL4, CCL3 and CXCL11) that are strongly correlated with each other. Re-examination of Figure 3.5 shows that there was no dominant cytokine grouping; in the ‘purple’ hyper-inflammatory cluster there was consistent elevation of all the measured protein biomarkers. Consideration of clusters determined by the relationships between measured cytokines and not patient samples would have failed to identify the distinct subsets of patients we have shown here.

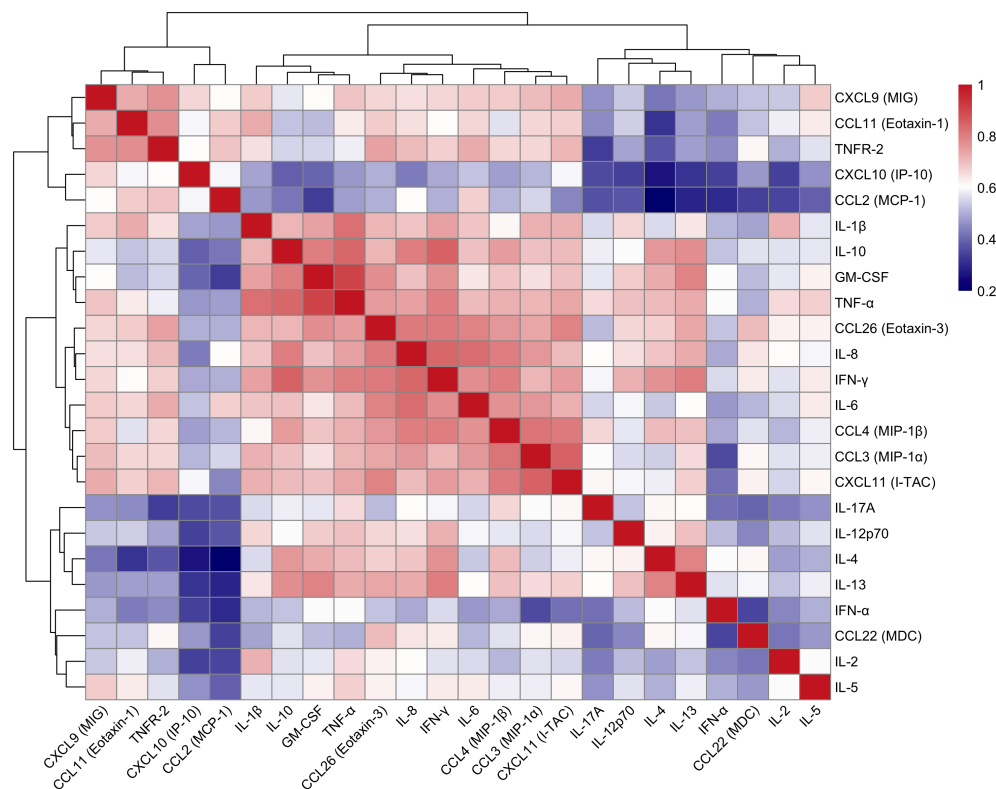


Fig. 3.13 Heatmap showing the correlations between protein biomarkers in patients with sepsis from the GAINs study. Protein biomarkers that are consistently correlated together are grouped into hierarchical clusters using the Ward linkage method. Although this method identifies groups of protein biomarkers that have consistent correlations it does not recognise the heterogeneity between patients at a sample level. The colours of cells in this heatmap are determined by the value of Pearson's correlation coefficient (r).

At this stage of the analysis, the clusters could have been named based on the predominant cytokine signatures for each one as demonstrated in the heatmaps and scaled linear plots. However, these labels would not have attributed any mechanisms as to why these patterns of cytokine release were observed. Further mechanistic characterisation of clusters was required before aligning clusters with clinical variables and patient outcomes. The gene expression data from the microarray experiments was analysed to determine the important mechanisms distinguishing each cluster.

3.4 Analysis of microarray data

Clustering using protein biomarkers values in Section 3.2 demonstrated that there were different immune profiles of acute illness in sepsis, severe influenza and possibly ARDS. These immune profiles were further characterised using whole blood-derived gene expression experiments. Given that the predominant cell types in the blood that are actively transcribing genes are immune cells, data from these two domains might be expected to complement each other.

Analysis of the relative expression of genes between patients with subtypes of their parent syndromes may identify the immune processes or other mechanism responsible for each subtype. Differential gene expression analysis could have been conducted using a variety of methods. This section describes moderated t-statistic and co-expression network-based methods to find groups of genes that had varying expression levels in these patients.

3.4.1 Quality control and pre-processing

The GAinS and MOSAIC researchers both used the same version of IlluminaTM microarray (Illumina Inc., San Diego, CA, USA) to quantify gene expression levels. The gene probe identifiers from each of the arrays were therefore consistently labelled, and no additional steps were required to address discrepancies in gene labels. In the GAinS study, the investigators measured gene expression using four microarray experiments. Of the 46,358 available probes only 22,972 were consistent across all four of these arrays. The inconsistency of available probes between different microarray chips served to highlight a limitation of microarray experiments; even when the same chip manufacturer and product version is used, there is considerable variation between each chip *before* the samples are subjected to any experimental or environmental sources of variation. The MOSAIC study researchers carried out the gene expression analysis on a single array experiment. Only the gene probe identifiers common to the GAinS microarray probes were used to ensure that all results downstream might be comparable.

After quantile and robust spline regression normalisation, probe intensities were plotted made to ensure normalisation steps were adequate (Figure 3.14). Multi-dimensional scaling (MDS) plots were used to assess the adequacy of batch-effect correction. MDS is another approach to visualising high-dimensional data, similar to PCA, but it uses rotation and location, not variance, to project points onto new axes. These plots showed that this method appeared to adjust the results from each of the GAinS microarray experiments appropriately (Figure

3.15). Pre-processing of the MOSAIC microarray results was conducted in the same way but without the need for additional batch correction steps.

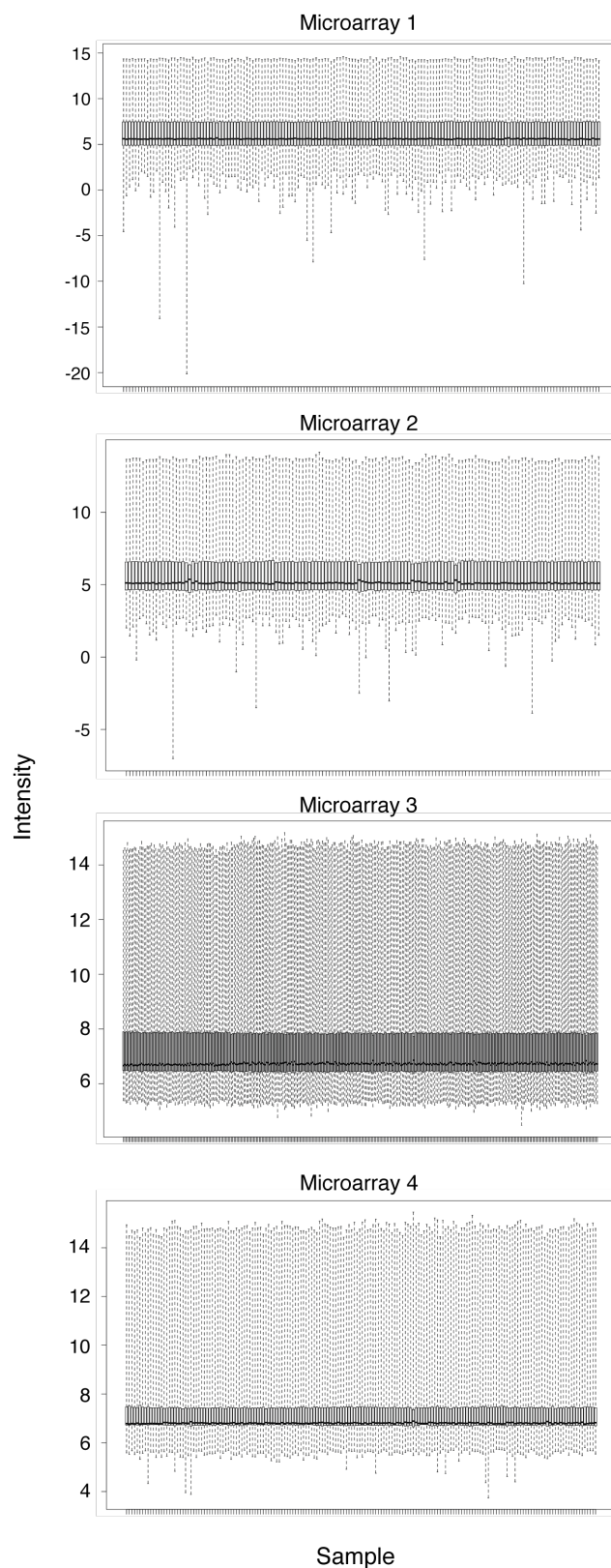


Fig. 3.14 Boxplots showing the intensity variation for all probes across each sample and each microarray experiment in the GAINs study. Although the intensities have been normalised for all samples in a given microarray, there remained considerable intensity variation between experiments which required adjustment with a batch correction method.

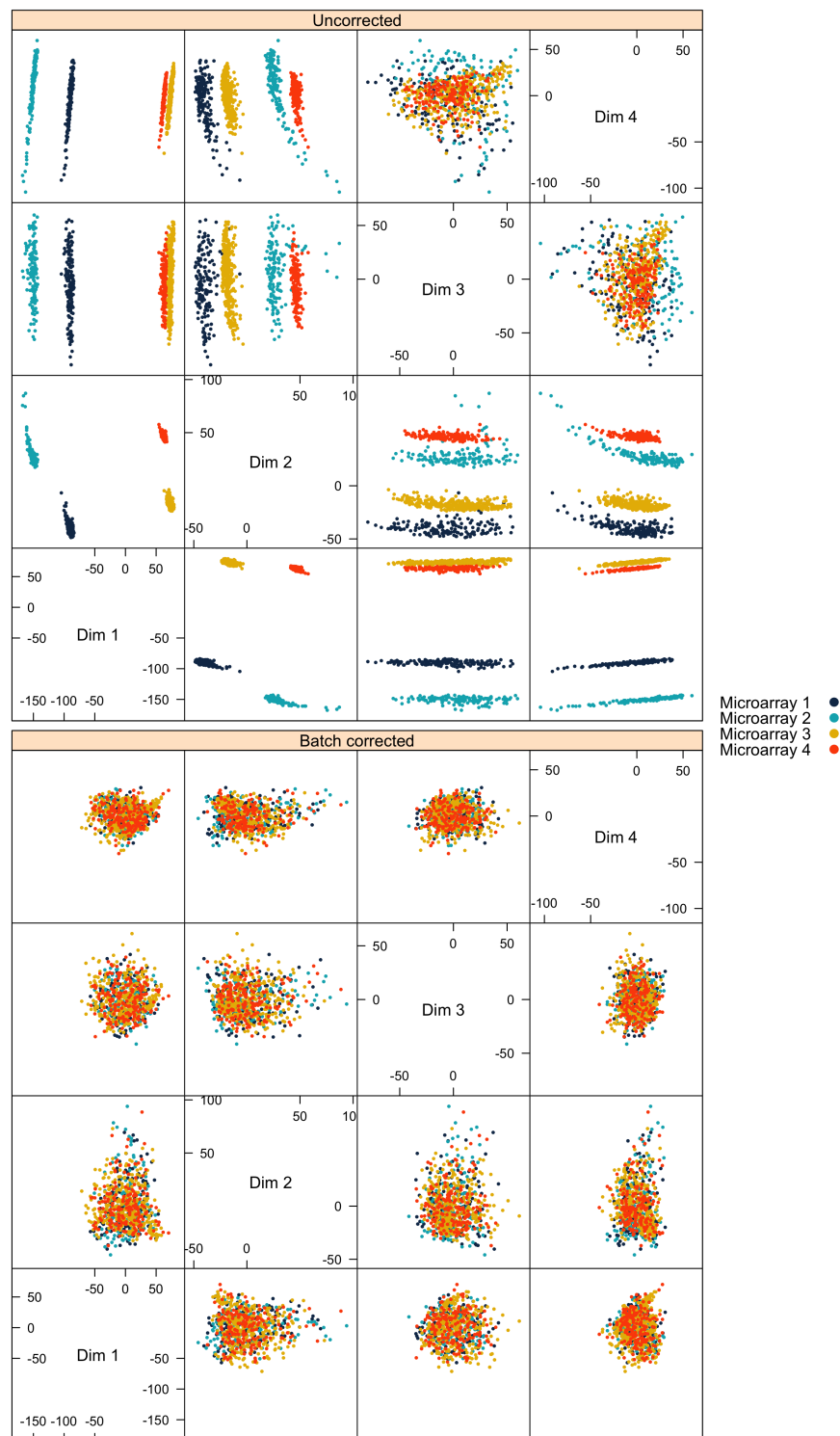


Fig. 3.15 Multidimensional scaling scatter plots showing the first four dimensions of the data from each of the GAINs microarray experiments. Each subplot is a pairwise comparison of two projected dimensions from MDS. Each point represents a patient sample encompassing all the values from 22,129 genes. Points are clearly grouped based on the microarray experiment they belong to in the uncorrected subplots, but this was mitigated by using batch effect correction as shown in the lower, batch corrected plot. Dim: MDS dimension

3.4.2 Differential gene expression analysis provides few insights in patients with sepsis and ARDS

Clinical variable information associated with each sample (metadata) was used to identify patients with features consistent with ARDS. The criteria for ARDS was based on sample $\text{PaO}_2\text{-FiO}_2$ ratio, PEEP levels and chest radiograph appearances as per the Berlin ARDS definition.⁴ Differential gene expression analysis between patients with and without sepsis-associated ARDS was performed using the *limma* library.

Surprisingly, there were no differentially expressed genes with $\text{FDR} < 0.05$ between patients with and without ARDS from the GAINs study. Although there were 462 genes that were differentially expressed with $p < 0.05$, none of these p values remained significant after multiple testing correction.

The volcano plot in Figure 3.16 shows the relative distribution of transcripts expression levels between patients with and without ARDS, based on uncorrected p values and their relative \log_2 fold change in expression. The twenty genes with the lowest p values are labelled in red. The relative differential expression, with respect to fold change, is very modest. The largest \log_2 fold change was -0.28 (transcript probe for HLA-DRB5, labelled in blue) which represents a 53% reduction in expression levels. Although there were no genes that were achieved statistical significant, the shape of this plot was satisfactory and not overtly skewed in any one direction. This demonstrated that the upstream processing of the microarray data and batch correction methods were adequate.

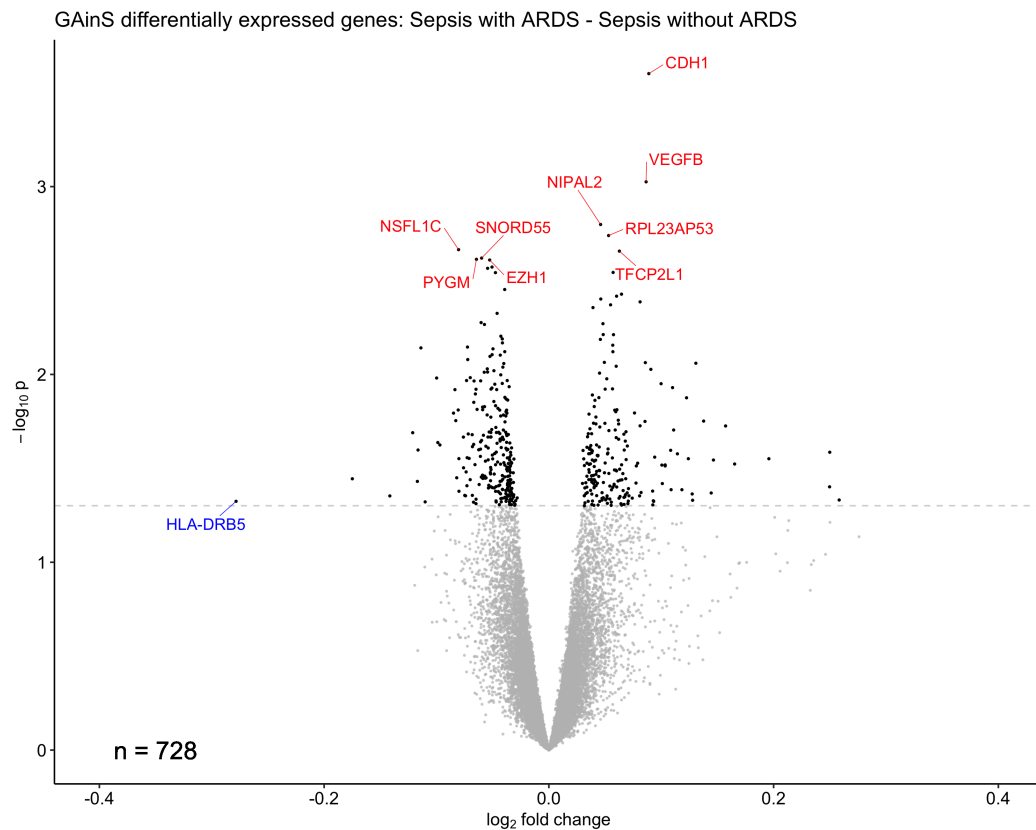


Fig. 3.16 Volcano plot showing the distribution of *unadjusted* p values and \log_2 fold changes for differentially expressed transcripts between patients with and without ARDS from the GAinS study. Each point is a gene transcript, with the black points representing gene transcripts with $p < 0.05$. The ten transcripts with lowest p values are highlighted in red, and transcript with largest magnitude in fold change is highlighted in blue.

3.5 Weighted gene co-expression network analysis of microarray results

Differential gene expression analysis between patients with and without sepsis-associated ARDS recruited to the GAINs study yielded no insights into the mechanisms underlying ARDS. For these reasons, a weighted gene co-expression network-based (WGCNA) approach was used to identify groups of co-expressed gene transcripts (gene modules). In order for the results from MOSAIC and GAINs studies to be comparable the gene probes that were common to all five microarray experiments (four from GAINs, one from MOSAIC), were used. 22,972 gene probes were common across all five microarray experiments.

3.5.1 WGCNA identifies gene modules in the microarray results from the GAINs and MOSAIC studies

The optimum soft power threshold values (β) were calculated for each microarray experiment individually and assessed graphically to identify the lowest value of beta that would achieve a within cluster sum of squares $R^2 > 0.8$. This threshold was consistent with a scale free network. For the microarray results from the GAINs study this value was equal to seven. For the microarray results from the MOSAIC study the optimum beta value was also equal to seven.

Construction of a scaled network for gene expression data from multiple microarray experiments required the use of the WGCNA *BlockwiseModules* function. This tool enabled identification of consistent modules across all four GAINs microarray experiments, avoiding the need to construct networks individually for each microarray experiment which would be difficult to compare directly. A minimum module size of 30 and dynamic cut height of 0.25 were used for determination of gene modules for the microarray results from both the GAINs and MOSAIC studies.

WGCNA produced a dendrogram which was organised into branches called gene modules. Each module was represented by a module eigengene (ME) which was the first principle component of each gene module. MEs were assigned numbers that were labelled as colours. The results were viewed as a dendrogram plot showing the heights (dissimilarity) of each module. A coloured bar beneath the plot showed the allocation of genes to different modules. These results along with the graphical determination of soft power thresholds are shown in Figure 3.17.

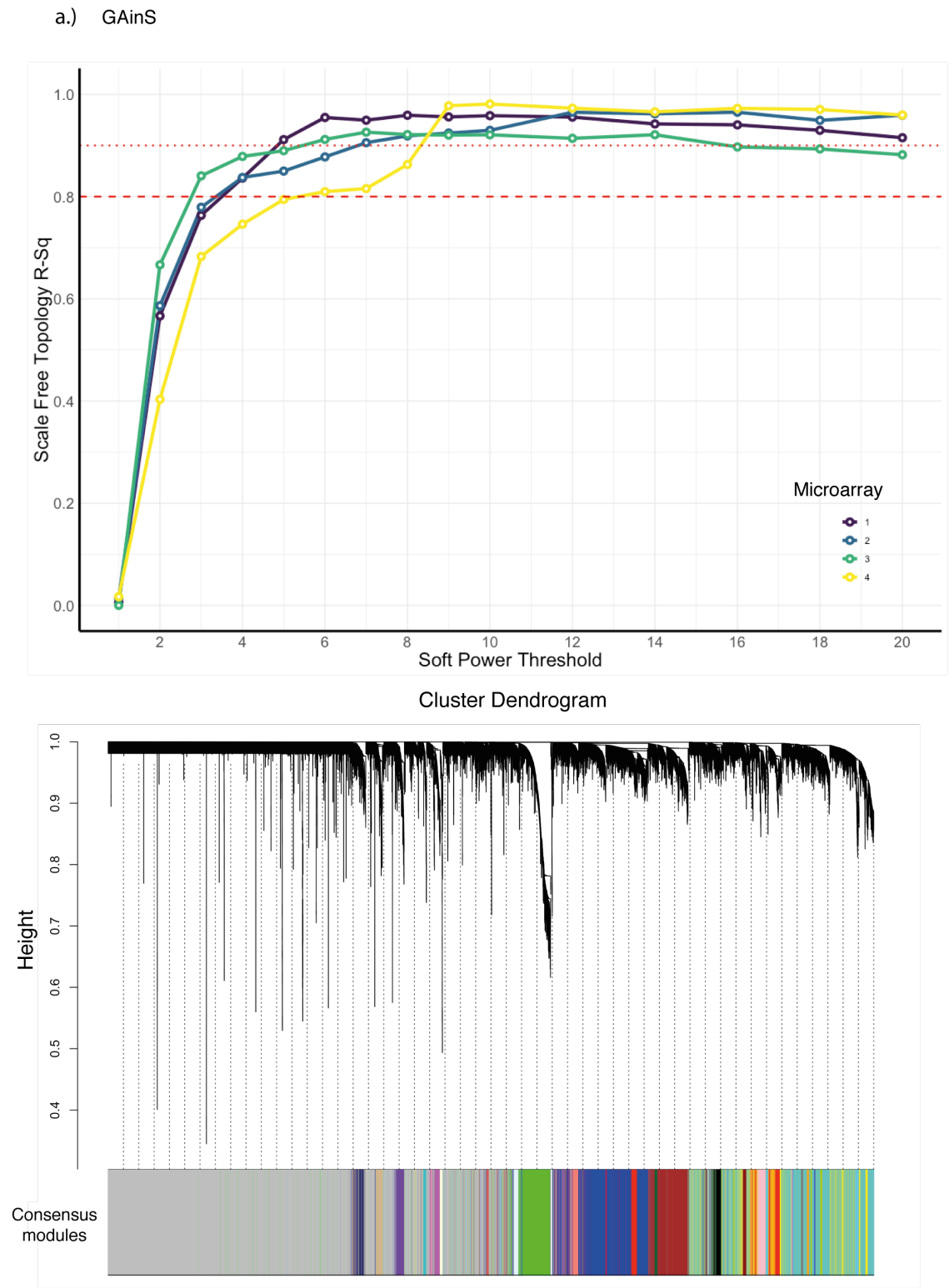


Fig. 3.17 (a) Full caption follows sub-figure (b)

b.) MOSAIC

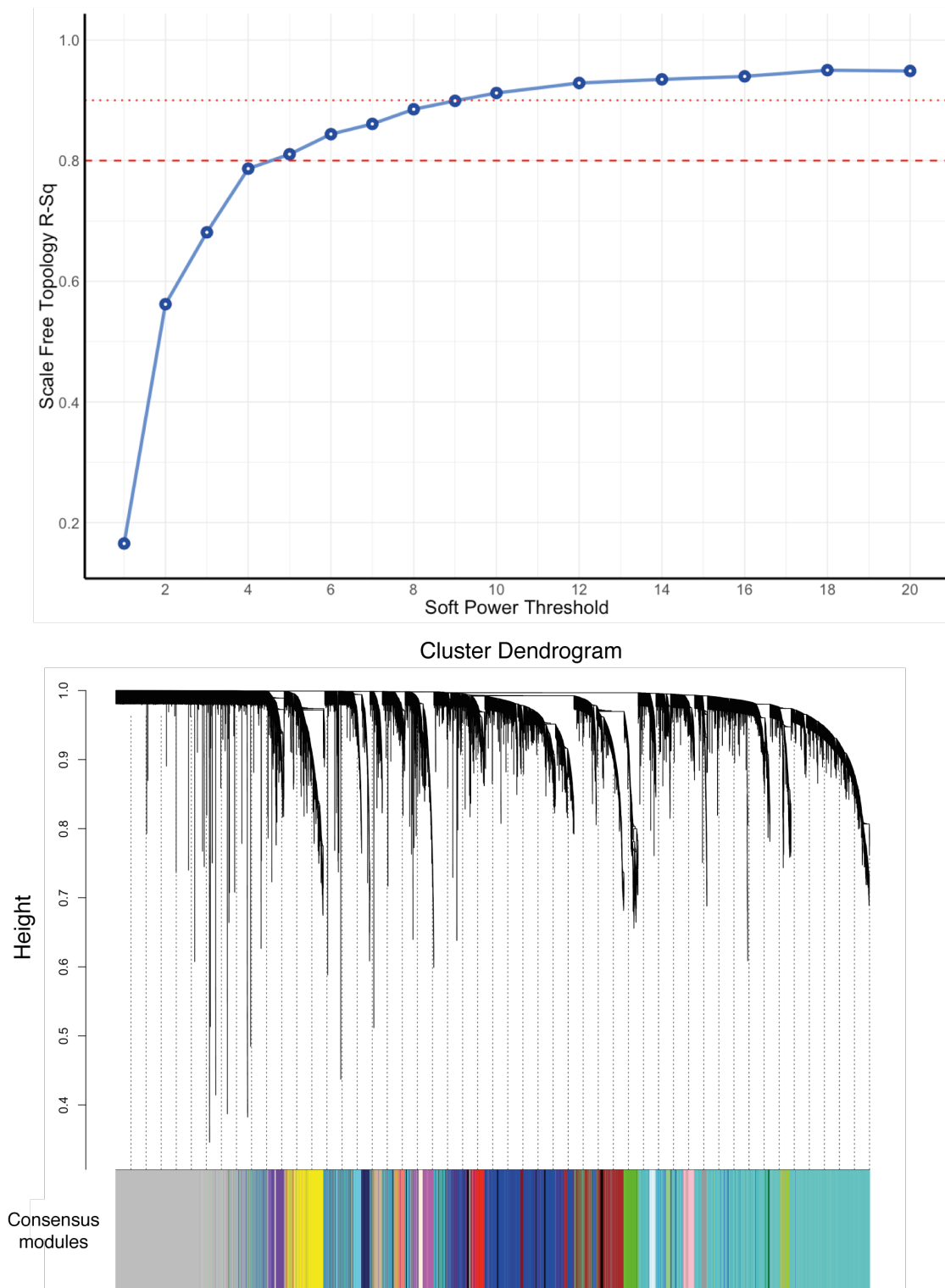
Fig. 3.17 (b) *Full caption on following page*

Fig. 3.17 Soft power threshold curves (upper) and cluster dendrograms (lower) for microarray results from the GAINs (a) and MOSAIC (b) studies. Dashed red lines represent R^2 thresholds of 0.9 and 0.8. A soft power threshold that has a scale free topology R^2 value greater than 0.8 is representative of a scale free network.

Cluster dendrograms show the distribution of gene modules assigned by using the *blockwiseModule* functions in the WGCNA R package. Consensus modules were assigned across all the GAINs microarray data using a soft power threshold equal to seven. Each colour corresponds to a family of highly connected transcripts that are branches of the cluster dendrogram. These have been isolated using a hybrid, dynamic cutting method developed by the WGCNA authors.

The grey band represents genes not assigned to a module. Note that colours and modules from each study do not correspond to the same gene modules between studies as these networks have been determined independently.

Genes that could not be assigned to a module were assigned to a ‘grey’ (ME0) module. This group represents the group of nodes in the network that are poorly connected or are pruned after application of the soft power threshold function.

The *consensusBlockwiseModules* function identified 26 modules in the GAINs microarray results. The WGCNA *BlockwiseModules* function identified 25 modules in the MOSAIC microarray results. Lists of genes from each module were submitted to the metascape tool (metascape.org, version 3.5) for statistical over-representation to determine the biological pathways represented by each gene module. These were cross checked by submitting the same gene lists to the enrichR ontology database (<https://maayanlab.cloud/Enrichr/>, accessed November 2020) tool.^{131,132} The enrichment results for each module are presented in Tables 3.6 and 3.7 for the microarray results from the GAINs and MOSAIC studies respectively.

Module name	Size	Principal gene ontology	Reference	Hyper-geometric <i>p</i> value
Turquoise	2820	Viral mRNA Translation	R-HSA-192823	1.90E-41
Blue	2183	Toll receptor signalling pathway	GO:0002224	2.23E-03
Brown	1497	NIK/NF-kappaB signalling	GO:0038061	1.93E-02
Yellow	1128	Nuclear mRNA surveillance	GO:0071028	9.41E-03
Green	1082	ATP synthesis, proton transport	GO:0042776	4.09E-02
Red	748	Maturation of LSU-rRNA	GO:0000470	8.01E-03
Black	321	Neutrophil activation	GO:0002283	7.90E-06
Pink	266	Thymic T cell selection	GO:0045061	9.10E-07
Magenta	247	Mitotic spindle assembly	GO:0051256	1.99E-02
Purple	245	Antigen presentation via MHC Ib	GO:0002476	3.37E-04
Green yellow	240	Stress-activated MAPK cascade	GO:0051403	2.70E-02
Tan	217	Protoporphyrinogen IX process	GO:0046501	3.11E-02
Salmon	217	Intracellular transport	GO:0046907	4.57E-02
Cyan	209	SRP membrane targeting	GO:0006614	2.00E-62
Midnight blue	205	Platelet aggregation	GO:0070527	1.15E-02
Light cyan	144	Regulation of transcription factors in hypoxia	GO:0061419	4.53E-02
Grey60	141	Platelet degranulation	GO:0002576	6.58E-03
Light green	105	Regulation of protein phosphorylation	GO:0001934	2.01E-03
Royal blue	104	No significant pathway	-	-
Light yellow	104	Neutrophil mediated killing of bacterium	GO:0070944	1.86E-02
Dark red	98	Cell adhesion	GO:0045785	1.36E-02
Dark green	93	Regulation of organelle assembly	GO:1902115	1.05E-02
Dark turquoise	89	IRE-1 mediated protein response	GO:0036498	7.96E-03
Dark grey	77	Neutrophil degranulation	GO:0043312	1.93E-08
Orange	63	Antigen processing and presentation	GO:0019882	1.26E-02
Dark orange	39	No significant pathway	-	-
White	32	No significant pathway	-	-
Grey	10258	Background genes	-	-

Table 3.6 Enrichment results for gene modules identified by WGCNA in the GAINs microarray results. Hyper-geometric *p* values were corrected using the Benjamini-Hochberg method.

Module colour	Size	Principal ontology / pathway	Reference	Hyper-geometric <i>p</i> value
Turquoise	6912	Translation	R-HSA-72766	2.66E-55
Blue	3310	Neutrophil degranulation	GO:0043312	1.56E-09
Brown	2008	RHO GTPases activate WASPs and WAVEs	R-HSA-5663213	1.04E-07
Yellow	971	Heme biosynthesis	R-HSA-189451	1.68E-02
Green	576	No significant pathway	-	-
Red	497	Toll Like Receptor 4 (TLR4) cascade	R-HSA-166016	3.04E-02
Black	463	Neutrophil degranulation	GO:0043312	2.58E-12
Pink	449	Cytoplasmic sequestering of NF-kappaB	GO:0007253	4.13E-02
Purple	418	Platelet degranulation	GO:0002576	9.70E-13
Magenta	431	Interferon alpha/beta signalling	R-HSA-909733	5.28E-27
Green yellow	314	L13a-mediated translational silencing of Ceruloplasmin expression	R-HSA-156827	5.19E-40
Salmon	280	No significant pathway	-	-
Cyan	252	G1/S-specific transcription	R-HSA-69205	3.90E-19
Midnight blue	251	Antimicrobial humoral response	GO:0019730	2.05E-09
Light cyan	197	Regulation of IL-2 production	GO:0032663	2.23E-02
Grey60	165	IRE1-mediated unfolded protein response	GO:0036498	5.26E-09
Light yellow	159	Interferon Signaling	R-HSA-913531	1.94E-05
Light green	159	Regulation of B-cell receptor signalling pathway	GO:0050855	2.77E-03
Royal blue	126	Regulation of expression of SLITs and ROBOs	R-HSA-9010553R	6.30E-04
Dark red	125	No significant pathway	-	-
Dark green	124	Cellular defense response	GO:0006968	3.01E-07
Dark turquoise	110	No significant pathway	-	-
Dark grey	86	Neutrophil degranulation	GO:0043312	2.19E-03
Orange	81	Eukaryotic translation elongation	R-HSA-156842	1.20E-04
Dark orange	60	No significant pathway	-	-
White	43	Innate immune response in mucosa	GO:0002227	4.33E-03
Sky blue	37	No significant pathway	-	-
Grey	4028	Background genes	-	-

Table 3.7 Enrichment results for gene modules identified by WGCNA in the MOSAIC microarray results. Hyper-geometric *p* values were corrected using the Benjamini-Hochberg method.

3.5.2 Module adjacency identifies closely related modules

The output from WGCNA included information that related each transcript to each module. The relative adjacency between each module could then be calculated using the euclidean distance metric and this was plotted on a dendrogram using agglomerative hierarchical clustering with average linkage. The module adjacency dendrograms for the the GAINs and MOSAIC studies are shown in Figure 3.18.

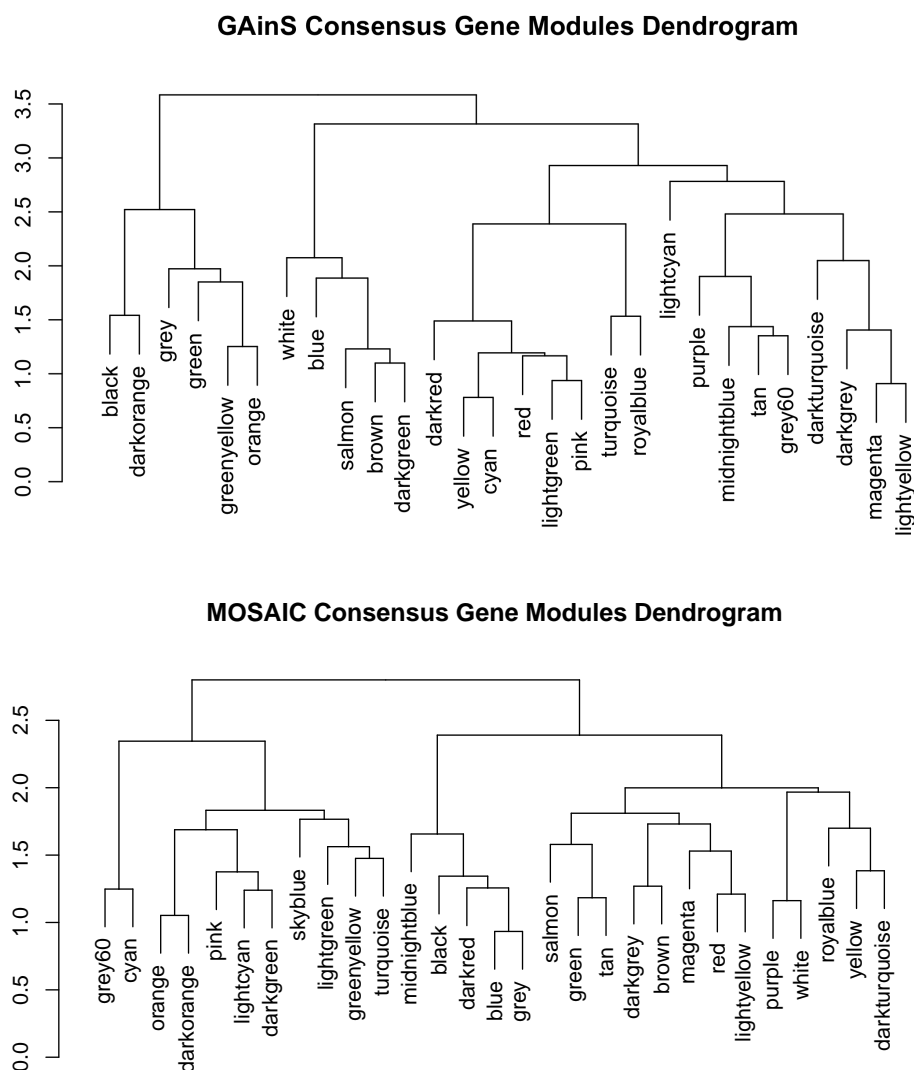


Fig. 3.18 Hierarchical clustering dendrograms of the distance between gene modules identified by WGCNA from the GAINs and MOSAIC microarray results. Modules that are closer together on this dendrogram are more similar even though each module contains different transcripts. Note that the module labels (as colours here) are arbitrarily named and do not imply concordance between networks derived from different data sources.

3.5.3 There is no significant correlation between gene modules and traits in patients recruited to the GAINs study

Pearson's correlation coefficient was calculated for clinical variables and diagnostic labels with gene modules. p values were adjusted using the Benjamini-Hochberg method and the results were visualised using a heatmap. Figure 3.19 shows the correlation coefficients and associated adjusted p values in each cell. This lack of correlation is consistent with the heterogeneity of patients with sepsis, even amongst those admitted to intensive care. It also serves to demonstrate how far removed biological processes are, at the gene expression level, from clinical measurements and clinical features in these patients. Of additional importance for this thesis, is the lack of any clear associations of gene modules with ARDS or any of the features relevant to ARDS (FiO_2 , $\text{PaO}_2\text{-FiO}_2$ ratio, PaCO_2 , severity grade of ARDS).

Module-Trait Relationships: GAIN Study Microarray Consensus Modules



Fig. 3.19 Heatmap showing the correlations between gene modules and clinical variables. The vertical axes labels are the gene modules, named by colour, with the number of genes contained in each module denoted in parentheses. Each cell contains the value of Pearson's r and the adjusted p value below in parentheses. The colour bar scale is for correlation coefficient values (r) used to colour the cells. There were no statistically significant correlations between gene modules and clinical variable or diagnosis labels.

CAP: community acquired pneumonia, FP: faeculent peritonitis, FiO2: fraction of inspired oxygen, PaO2: partial arterial pressure of O2, PaCO2: partial arterial pressure of carbon dioxide, AHRF: acute hypoxic respiratory failure, MAP: mean arterial pressure, HR: heart rate, WCC: white cell count

3.5.4 Correlation between clinical variables and gene modules identify plausible biological processes in patients recruited to the MO-SAIC study

Biological samples were temporally matched to patients' clinical variable measurements. The constituent components of the SOFA score were matched to these sampling times. Scores consistent with major organ dysfunction (SOFA 3 or 4 for each organ system) were dichotomised at this level instead of making comparisons at every ordinal level of the SOFA scale. The central nervous system SOFA score was not correlated with gene modules as this field was largely incomplete in the study database. Correlations between the clinical variables and gene module eigenvalues ('eigengenes') were calculated along with p values. p values were adjusted using the Benjamini-Hochberg method.

Figure 3.20 shows there are a number of gene modules that were significantly correlated (negatively and positively) with different patient characteristics. Of note is the 'midnight blue' module which positively correlated with: poor respiratory and cardiovascular SOFA scores, raised white cell and neutrophil counts and raised creatinine levels. Enrichment analysis of the transcripts contained in this module found the principal pathway associated with this module was 'antimicrobial humoral response' (GO:0019730, adjusted $p = 2.1 \times 10^{-9}$).

Other modules with strong positive correlations across multiple clinical features included :

- 'Dark red' module: correlated with an increased risk of hospital mortality, higher respiratory and cardiovascular SOFA scores. This module enriched for the term 'organic substance catabolic process' (GO:1901575, adjusted $p = 9.75 \times 10^{-3}$).
- 'Black' module: correlated with high respiratory, cardiovascular SOFA scores and raised creatinine. this module enriched for the term 'natural killer cell degranulation' (GO:0043320, adjusted $p = 0.023$)

Several modules were negatively correlated with high cardiovascular and high respiratory SOFA scores and to varying degrees, death, white cell and neutrophil counts:

- 'Dark green' module: 'cellular defense response' (GO:0006968, adjusted $p = 3.01 \times 10^{-7}$)
- 'Pink' module: 'cytoplasmic sequestering of NF-kappaB' (GO:0007254, adjusted $p = 0.04$)
- 'Orange' module: 'viral transcription' (GO:0019083, adjusted $p = 0.002$)

- ‘Dark orange’ module: which did not enrich for a pathway.

The ‘dark green’, ‘pink’ and ‘orange’ modules enrich for plausible pathways related to immune function or viral infection. For example, in the pink module the process of cytoplasmic sequestering of NF-kappaB refers to the transcription factor NFκB. This protein, if released from its inhibitory protein Iκb, migrates to the nucleus where it stimulates transcription of genes associated with the inflammatory response and other major cellular processes.¹⁴⁶ Sequestration of NFκb would therefore be associated with suppression of inflammation.

The correlations shown in Figure 3.20 should be interpreted with some caution. There is a lack concordance between some clinical features and gene modules: there should be consistent correlations between gene modules with platelet count and coagulation SOFA scores. Similarly, the systolic blood pressure and cardiovascular SOFA scores should have consistent correlations with gene modules. Other observed correlations were more consistent, for example the white cell and neutrophil count correlations with gene modules. Inconsistencies may have arisen due to incomplete and inaccurate clinical data annotations within the study database. The MOSAIC database attempted to record as many as 14,000 clinical variables on each patient, across multiple study sites. The MOSAIC study had a broad remit and was not focused on measurements related to critical care alone.

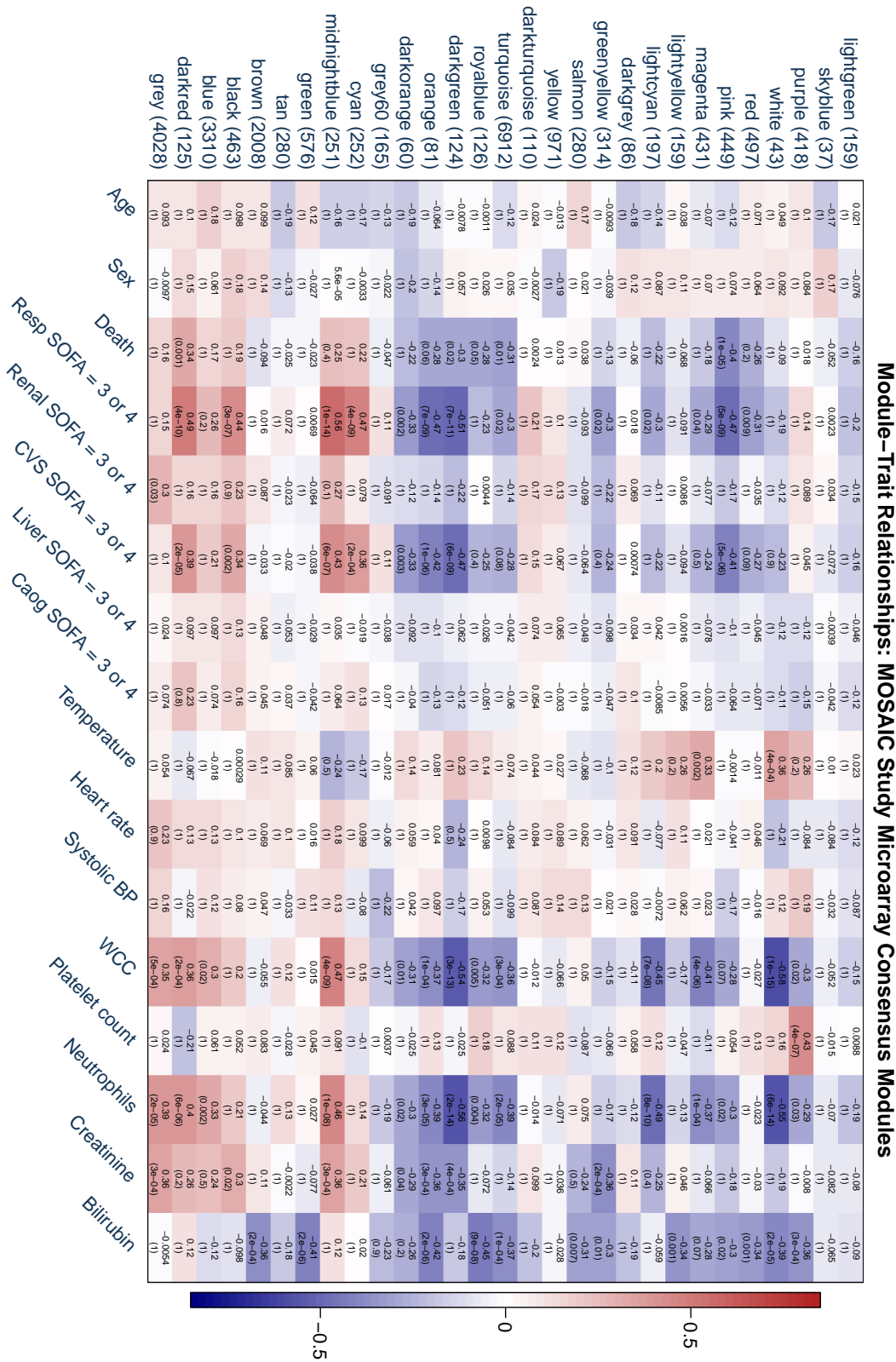


Fig. 3.20 Caption on following page

Fig. 3.20 Heatmap showing the correlations between gene modules identified by WGCNA and clinical variables of patients enrolled in the MOSAIC study. The vertical axes labels are the gene modules, named by colour, with the number of genes contained in each module denoted in parentheses. Each cell contains the value of Pearson's r and the adjusted p value below in parentheses. The colour bar scale is for correlation coefficient values (r) used to colour the cells.

There are a number of key correlations that are statistically significant: the 'midnight blue' gene module was significantly correlated with severe respiratory failure, cardiovascular dysfunction, white cell and neutrophil counts. There are inconsistent correlations here as well: the modules that bilirubin had significant correlations with were not consistently correlated with the "liver SOFA = 3 or 4" category. The hepatic component of the SOFA scores were determined by the bilirubin level. The same inconsistencies were observed between the platelet count and 'Coag SOFA = 3 or 4' columns. These were likely due to inconsistent database encoding of recorded clinical variables and missing data points in the database. Access to raw case report forms and source data were unavailable to resolve these inconsistencies. The death column implies hospital mortality.

SOFA: sequential organ failure score, CVS: cardiovascular system, Coag: coagulation, BP: blood pressure, WCC: white cell count.

3.6 Discussion of microarray and WGCNA results

The quality control and pre-processing steps demonstrated that the batch-effect variation in the microarray results from the GAINs study could be adjusted by using the ComBat method. However, these methods, in combination with the heterogeneity of patients with sepsis or ARDS, may have suppressed the possibility of identifying any gene expression-based signal associated with ARDS. There were no transcripts that were statistically different in expression between septic patients with and without ARDS (Figure 3.16). This was consistent with the hypothesis that direct comparisons between sepsis and ARDS are unlikely to be successful due to the heterogeneity of patients with either of these syndromes.

Surprisingly, none of the gene modules identified by WGCNA in the GAINs cohort were correlated with a diagnosis of ARDS or any other clinical feature (Figure 3.19). This was unexpected as the gene set enrichment analysis of the gene modules identified by WGCNA were related to plausible biological processes that might occur in patients with sepsis or ARDS. The failure to demonstrate a difference is unlikely to be due to an inadequate sample size as the GAINs study had 728 samples available for analysis which is larger than most other studies of this kind. The failure to demonstrate any strong correlations, or differentially expressed genes using the above methods suggests that a substantial source of heterogeneity amongst these patients remains unaccounted for.

The results from the MOSAIC study were more encouraging with respect to identification plausible biological mechanisms in patients with severe influenza infection. Dunning *et al* (2019) had already shown, by using clustering analysis, that severe cases of pandemic influenza were more likely to be associated with neutrophil-related ontology modules and mild cases were more associated with interferon-related ontology modules.¹¹⁰ Using WGCNA to determine gene modules in the MOSAIC samples identified several modules that were associated with patient features (Figure 3.20).

The ‘midnight blue’ module was significantly correlated with severe respiratory and cardiovascular dysfunction, raised white cell and neutrophil counts. The genes in this module enriched for the process ‘antimicrobial humoral response’ (GO:0019730, adjusted $p = 2.1 \times 10^{-9}$). This result suggested that immune mechanisms related to bacterial infection in these patients may have been activated. Secondary infection of patients with influenza pneumonia by bacteria is regarded as an important contributor to the mortality and morbidity of these patients. Due to the difficulties with obtaining samples from the lung in acutely unwell patients and the processes required for successful bacterial culture, it is often difficult to confirm the presence of secondary infection in patients clinically.

The positive correlation of this module with the patients traits: multi-organ dysfunction, raised white cells and neutrophil counts suggested that secondary infection may have been responsible for the clinical picture in patients with these gene expression profiles. It would be unwise to attribute the clinical features of these patients to bacterial infection based on this result alone as ontology analysis only considers the highest rated pathway to the exclusion of all others for a given set of genes. In addition, there may be overlap in immune responses to bacterial and viral infection.

Other key modules that were identified by these methods were the ‘dark red’ and ‘black’ modules, which both significantly correlated with organ dysfunction. Enrichment analysis of these modules attributed the genes contained within them to be associated ‘organic substance catabolic process’ and ‘natural killer cell degranulation’. Natural killer (NK) cells are a key immune cell involved in the innate immune response to viral infections and so their association with organ dysfunction is plausible. The catabolic processes associated with the ‘dark red’ module may be a reflection of the metabolic state during critical illness or a loss of homeostasis with respect to metabolic functions. Both the GAINs and MARS studies exploring the transcriptomic signatures in sepsis have reported the association of sepsis endotypes with genes that enrich for aberrant metabolic pathways.^{96,97}

Extending the integration of the gene expression and cytokine analysis might enable further characterisation of and insights into the processes that underlie the immune responses in ARDS and influenza infection which is the subject of the Chapter 4.

CHAPTER 4

Integration of protein biomarkers with transcriptomics

Section 3.4 demonstrated how gene modules identified in the GAINs study correlated poorly with clinical features. The same methods applied to the gene modules in the MOSAIC study identified some plausible mechanisms but these were inconsistent.

Protein biomarker-based clustering had shown relevant and distinct immune profiles in Section 3.2 and so it was a logical step to try to determine the gene expression signatures between different clusters. This chapter describes the results of standard analytical and novel methods to integrate gene expression with serum protein biomarkers.

4.1 Differential gene expression between clusters

4.1.1 Differential gene expression between protein biomarker-based clusters of patients with ARDS in the GAINs study

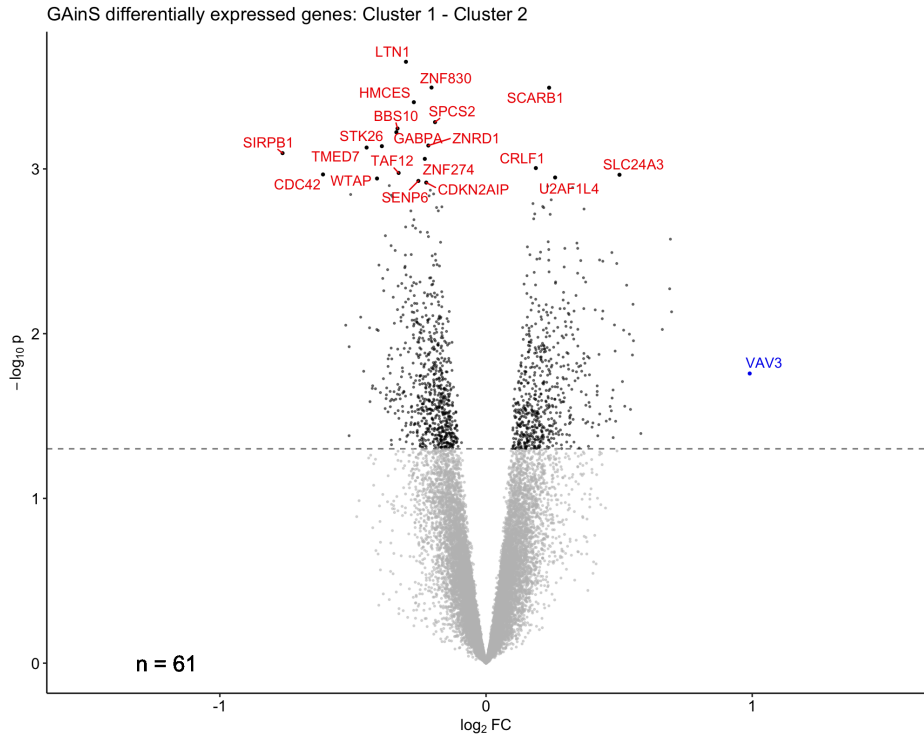
The microarray data from the GAINs study were filtered for samples that belonged to a cluster, determined by analysis of protein biomarker concentrations, and for patients with ARDS. After these filtering steps 71 samples remained.

A linear model was fitted to compare the differences between each of the three clusters using moderated t-tests with the empirical Bayes method described in the *limma* library.¹¹⁸ p values were adjusted using the Benjamini-Hochberg method. No genes in any of the pairwise comparisons between samples in each cluster, were statistically significant after correction of p values for multiple comparisons. Given the significant differences in protein

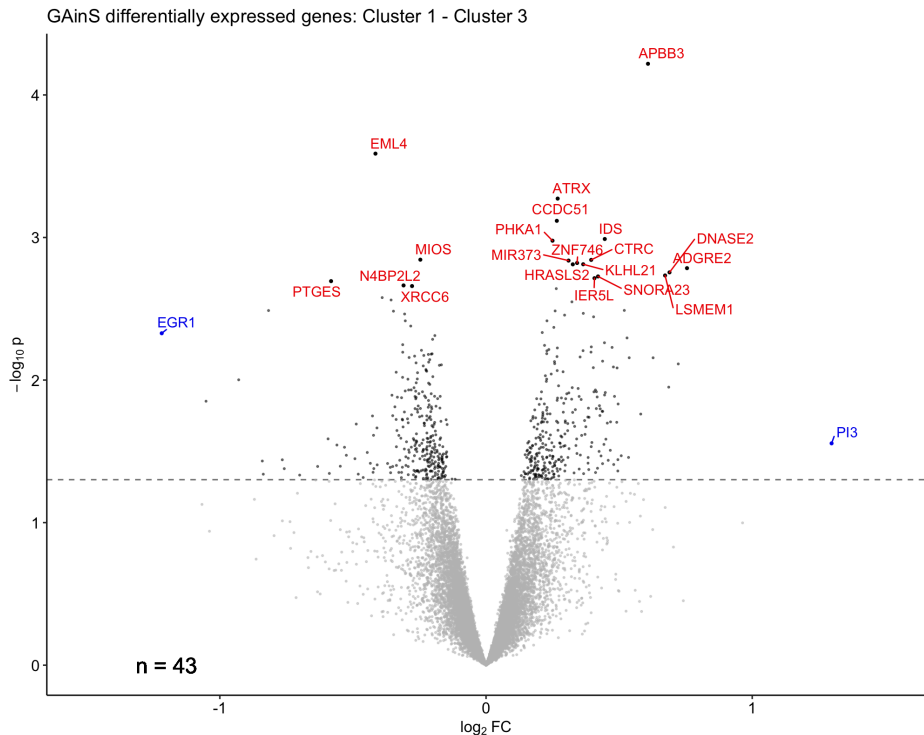
biomarker concentrations between samples from each cluster this was a surprising finding. Volcano plots for each comparison between clusters of patients who had sepsis and ARDS are shown in Figure 4.1. The vertical axes for each plot are the negative \log_{10} transform of the *unadjusted* p values. The transcripts above the $p < 0.05$ dashed lines on these plots were not significant after adjustment. These plots did, however, demonstrate that inter-cluster comparisons generated results that were distributed correctly following a differential gene expression analysis, and that the methods were satisfactory.

Upon review of the volcano plots in Figures 4.1b and 4.1c it was apparent that many of the labelled genes were common to both plots. There were 589 differentially expressed transcripts with $p < 0.05$ between the ‘green’ cluster (1) and ‘yellow’ cluster (3) (Figure 4.1b). There were 798 differentially expressed transcripts between ‘purple’ cluster (2) and ‘yellow’ cluster (3) with $p < 0.05$ (Figure 4.1c). Of the combined 1,396 transcripts with $p < 0.05$, 241 (17.3%) were in common. This was an unusually high value and probably reflected the low number of samples in the ‘yellow’ cluster (3) with ARDS ($n = 10$). Low samples numbers may have increased the influence of outlying sample on the fitted models.

To assess whether there was excess variance in the linear model comparing the gene expression levels between clusters, the residual standard deviation was compared with the mean average log expression. The *SAplot* function from the *limma* library identified no outlying values, which was consistent with no evidence of excess variance in these data.



(a)



(b)

Fig. 4.1 (Caption on following page.)

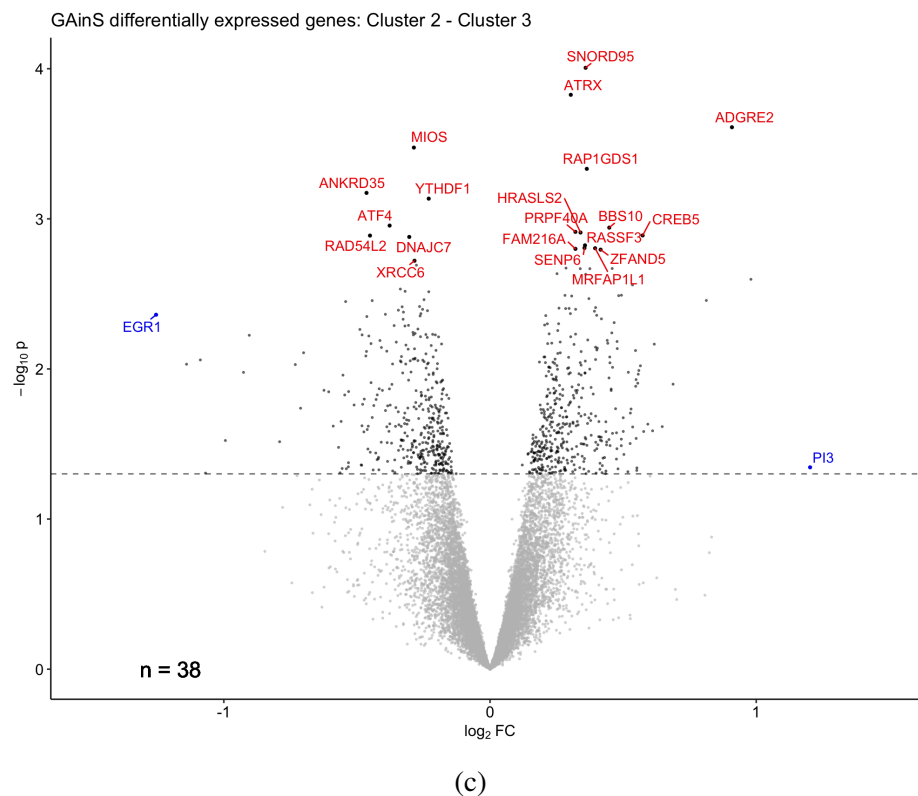


Figure 4.1 Volcano plots of differentially expressed transcripts between patients with ARDS in each cluster. None of these comparisons identified transcripts that were statistically significant after multiple comparison correction. The dashed lines represent an unadjusted $p = 0.05$. Points coloured black represent transcripts with *unadjusted* $p < 0.05$. Red labelled points are the 20 transcripts with the lowest p values and the blue labelled points are transcripts with the highest fold change.

4.1.2 Differential gene expression between protein biomarker-based clusters of patients with severe respiratory failure in the MO-SAIC study

The microarray results from the MOSAIC study were labelled and filtered for samples that belonged to a cluster, determined by analysis of immune mediator concentrations, and for patients with respiratory SOFA scores greater than two. After these filtering steps 100 samples remained.

A linear model was fitted to compare the differences between each of the three clusters using moderated t-tests with the empirical Bayes method in the *limma* R package. *p* values were adjusted using the Benjamini-Hochberg method. There were significant differences in the expression levels between the ‘blue’ (1) and ‘grey’ (2) clusters, and between the clusters ‘grey’ (2) and ‘red’ (3) clusters (Figure 4.2a and c). However, for the comparison between ‘blue’ (1) and ‘red’ (3) clusters only four transcripts were found to have an adjusted $p < 0.05$ (Figure 4.2b).

Differential gene expression results are often presented as a comparison between an experimental condition and control. The analysis may show that some genes may be up-regulated (\log_2 fold change > 0) whilst other genes be down-regulated (\log_2 fold change < 0) compared with control samples. Generally, all the differentially expressed genes, regardless of their directional changes, are submitted for enrichment analysis to identify the biological processes related to these genes. The purpose of this project was to delineate the differences between endotypes of critical care syndromes, and so comparisons were made between difference immunological states defined by the protein biomarker clusters.

Directional changes in gene expression were therefore considered as differentiating one cluster from the other. For example, if gene *A* had a greater fold change in cluster X compared with cluster Y, then gene *A* would be considered to have a higher relative expression in cluster X samples compared with cluster Y samples. Gene *A* would therefore help to delineate the processes taking place in that cluster X as it was relatively up-regulated in these samples compared with cluster Y. For this analysis, significantly expressed genes with a \log_2 fold change greater than zero were considered to have increased relative expression in a given cluster, whilst gene with \log_2 fold change lower than zero were considered to have increased relative expression in the other comparative cluster.

Three transcripts in particular were found to be down-regulated with respect to MOSAIC cluster 2 (grey): *CD177*, *RETN* and *ZDHHC19* (Figure 4.2a and c). *CD177* and *RETN*

are associated with immunological roles. CD177 is expressed on neutrophil cell surfaces and plays a role in activation and transmigration by interacting with $\beta 2$ integrins and platelet endothelial cell adhesion molecule 1 (PECAM1).¹⁴⁷ The *RETN* gene codes for the protein resistin which is associated with neutrophil degranulation and other immune functions.¹⁴⁸

Statistically significant genes with same directional change in expression (positive or negative \log_2 fold change), were submitted for enrichment analysis to determine the biological processes responsible for the differences between clusters. The number of transcripts that belonged to a given process and their adjusted p values are shown in the bubble diagrams below each volcano plot in Figure 4.2. The bubbles to the left of the zero $-\log_{10}$ FDR line represent the enriched pathway for transcripts with \log_2 fold change below zero in the volcano plot above it. Similarly the bubbles to the right of the zero $-\log_{10}$ FDR line represent enrichment of transcripts with a positive \log_2 fold change in the associated volcano plot.

The bubble plot in Figure 4.2a shows that the transcripts with \log_2 fold change less than zero enriched for biological pathways associated with ‘lymphocyte activation’ (GO:0046649) and ‘adaptive immune response’ (GO:0002250). This was consistent with their immune mediator profiles as ‘grey’ cluster (2) was associated with raised levels of CCL5 (RANTES), CC17 (TARC) and CCL22 (MDC) all of which are associated with lymphocyte activation and recruitment (Figure 3.9).¹⁴⁹ Similarly, the transcripts with \log_2 fold change greater than zero were enriched for ‘neutrophil degranulation’ (GO:0043312). The higher concentrations of TNF- α , IL-6 and IL-8 in ‘red’ cluster (3) samples, were consistent with these observed associations (Figure 3.9).

Figure 4.2a suggested that the differences between samples in the ‘grey’ and ‘red’ clusters might be explained by the relative lymphocyte and neutrophil counts in each cluster. In which case gene expression would have been no better at distinguishing the clusters than a routine full blood count test. Figure 4.3 shows that there were no differences in the relative lymphocyte counts between clusters and that the ‘blue’ MOSAIC cluster had significantly higher neutrophil counts compared with the ‘grey’ cluster. There were no differences in the neutrophil counts between samples in the ‘blue’ and ‘red’ clusters.

Figure 4.2c shows that the transcripts with \log_2 fold change less than zero enriched for mechanisms relating to ‘haemostasis’ (R-HSA-109582) and that this process was relatively more important in the ‘grey’ MOSAIC cluster (2) compared with ‘blue’ MOSAIC cluster (1). Differential gene expression analysis did not identify any pathways or processes associated with up-regulated genes in samples from the ‘blue’ cluster (1).

There were only 4 transcripts that were differentially expressed between the ‘blue’ (1) and ‘red’ (3) MOSAIC clusters (Figure 4.2b). This was unexpected and the shape of the volcano plot suggested either there were no differences between the two clusters or that this method was unable to determine any differences in gene expression between these two groups.

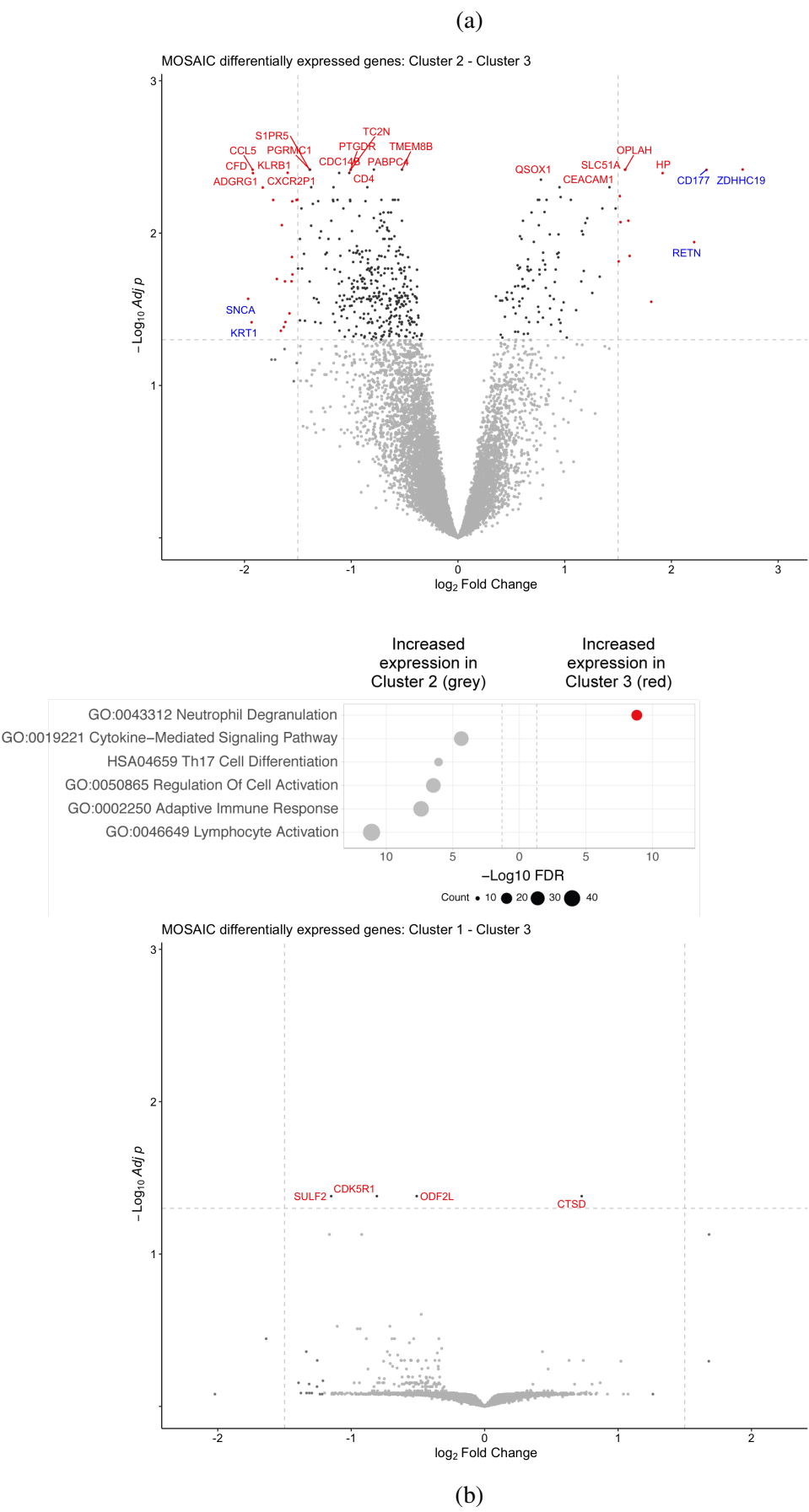


Fig. 4.2 (Caption on following page.)

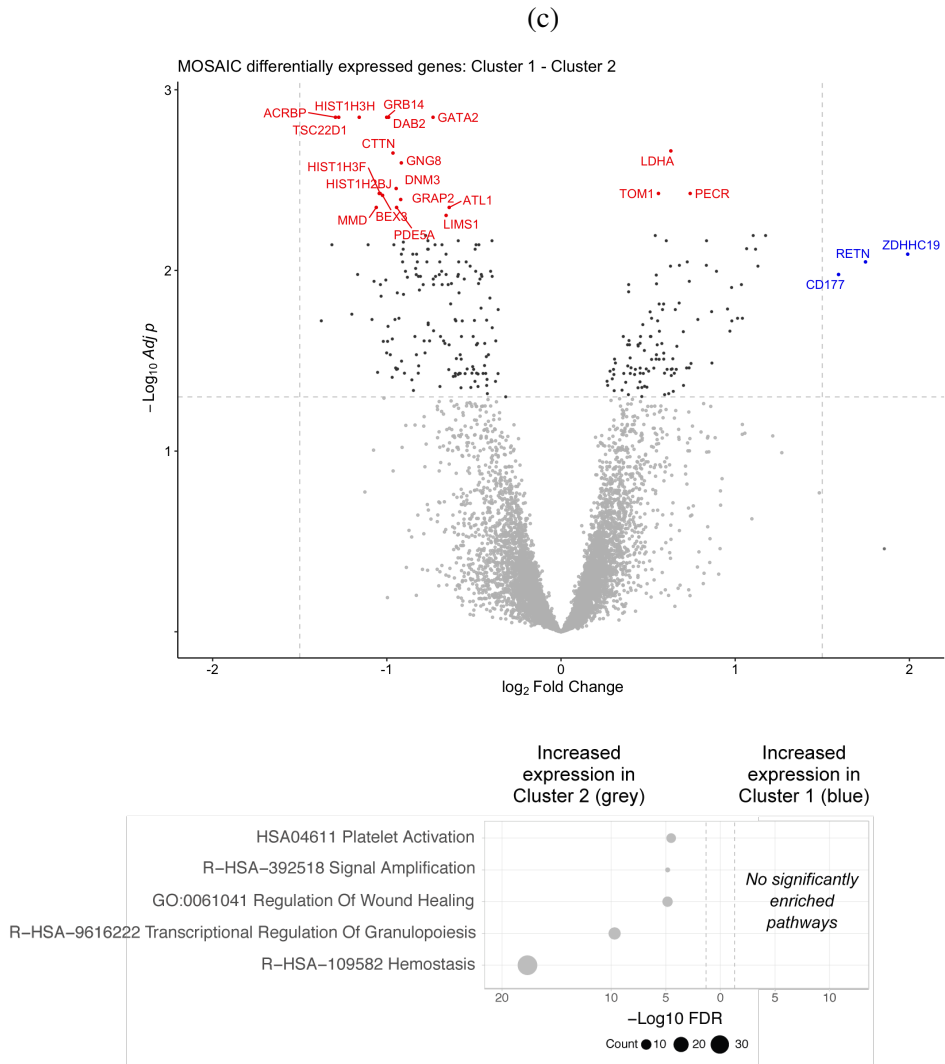


Figure 4.2 Volcano plots of differentially expressed genes between patients with severe respiratory failure ($\text{rSOFA} \geq 3$) in each cluster from the MOSAIC study. The vertical axes are for the adjusted p values. Red labelled points are the 20 transcripts with the lowest p values and the blue labelled points are transcripts with the largest fold change. The horizontal dashed line is the adjusted $p < 0.05$ threshold. The vertical dashed lines are the \log_2 fold change values of -1.5 and 1.5 respectively. Gene lists with significant adjusted p values from each ‘half’ of the volcano plot were analysed using Metascape (<http://metascape.org>, version 3.5) and enrichR (<https://maayanlab.cloud/Enrichr/>). The number of genes that enriched for a given biological process and their relative adjusted p values ($-\log_{10}$ FDR) are shown below each volcano plot as a bubble chart. The size of each bubble denotes the number of genes associated with a process. Note that the horizontal axis on the bubble charts is positive in both directions. No enrichment diagram is presented for (b) as the four transcripts with adjusted $p < 0.05$ did not enrich for a biological process.

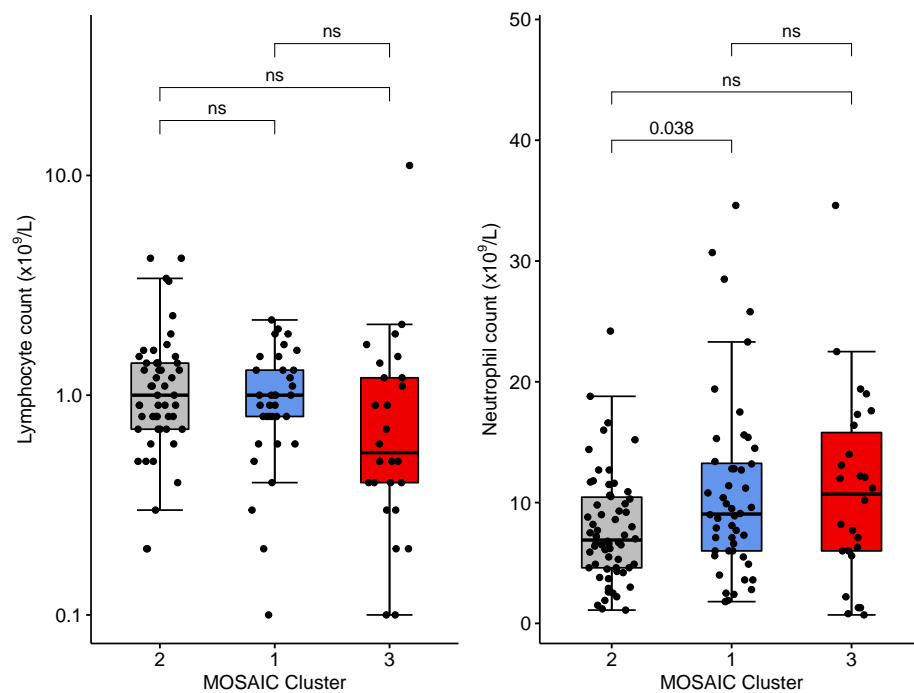


Fig. 4.3 Boxplots showing the relative levels of lymphocytes and neutrophils in samples from each of the MOSAIC clusters using paired full blood count results at sampling times T1 and T2. There were no differences in cell counts for these two cell types between the 'red' and 'blue' clusters. There were significantly more neutrophils in the 'blue' MOSAIC cluster compared with the 'grey' cluster. Comparisons were made using ANOVA with Tukey's post hoc test. NS: not significant.

4.2 Gene module correlation analysis

4.2.1 There are no gene modules identified in the GAINs study that significantly correlated with protein biomarker clusters

The module eigengene values for each patient sample were correlated with membership of each cluster using Pearson’s method. Figure 4.4 shows there were no significant correlations.

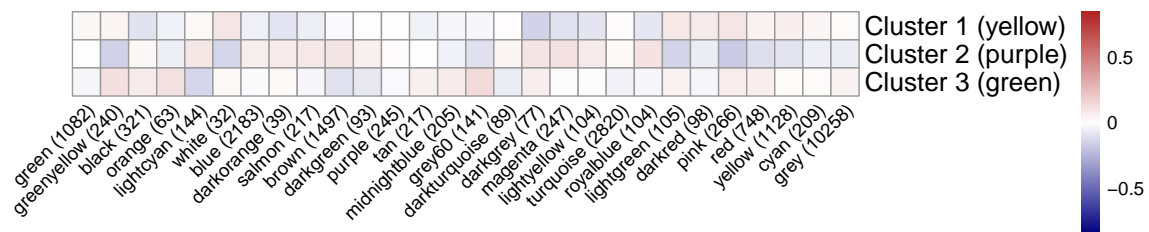


Fig. 4.4 Heatmap showing the correlations (using Pearson’s r) between gene modules identified by WGCNA and the clusters determined by the serum cytokine concentrations for patients in the GAINs study. The colour bar scale is for correlation coefficient values (r) used to colour the cells. The calculated correlation coefficients between cluster and gene modules were higher in magnitude compared with the correlations between clinical variable and gene modules shown in Figure 3.19. None of the correlations shown in either heatmap achieved statistical significance.

4.2.2 Correlation between protein biomarkers clusters and gene modules identified important mechanisms in the ‘red’ MOSAIC cluster

The correlation heatmap between cytokine clusters and gene modules is shown in Figure 4.5. The ‘green yellow’ gene module showed a significant correlation. The ‘red’ cluster (3) was positively correlated with this module ($r = 0.37$, adjusted $p = 0.01$). The same module was negatively correlated with the ‘grey’ cluster (2) but this was not significant after correction for multiple comparisons ($r = -0.27$, adjusted $p = 0.25$).

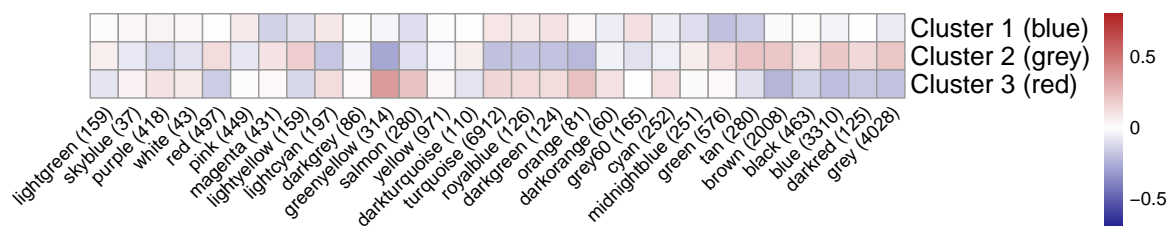


Fig. 4.5 Heatmap showing the correlations (using Pearson’s r) between gene modules identified by WGCNA and the clusters determined by the serum cytokine concentrations for patients in the MOSAIC study. The colour bar scale is for correlation coefficient values (r) used to colour the cells. There was a significant positive correlation between the “green yellow” module and cluster 3 (red) ($r = 0.36$, adjusted $p = 0.01$). The same module was negatively correlated with cluster 2 (grey) ($r = -0.27$, $p = 0.006$) but this was not significant after adjustment (adjusted $p = 0.25$).

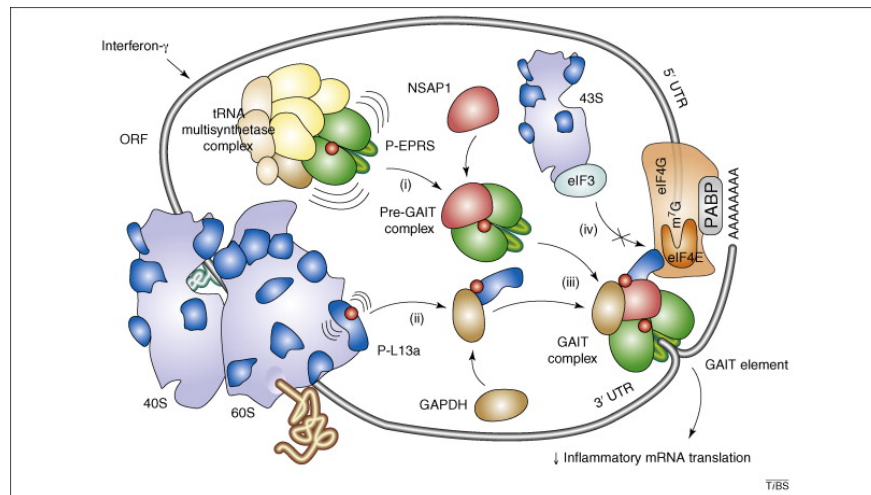


Fig. 4.6 Cartoon of the gamma-interferon inhibition of translation (GAIT) mechanism, which was positively correlated with the 'red' MOSAIC cluster.

Taken, with permission, from Figure 2 of Rupak Mukhopadhyay, Jie Jia, Abul Arif, Partho Sarothi Ray and Paul L. Fox.

The GAIT system: a gatekeeper of inflammatory gene expression.

Trends in Biochemical Sciences 2009 34(7):324-31.

doi: 10.1016/j.tibs.2009.03.004.

Copyright Elsevier Ltd. 2020

The licence for use of this figure is available in Appendix H

Enrichment of the genes from the 'green yellow' module showed that it was associated with the process 'L13a-mediated translational silencing of caeruloplasmin' (R-HSA-156827, $p \ll 0.0001$). This process is also known as the gamma-interferon associated inhibition of translation (GAIT) mechanism. The GAIT mechanism is a late response to IFN- γ signalling which causes the arrest of translation of caeruloplasmin mRNA. Caeruloplasmin is a copper ion binding protein that is released in the acute phase of the inflammatory response. The mRNA for this protein has a particular 3-prime untranslated region (3' UTR) that has a circular structure that is called a GAIT motif. This motif can be recognised by a number of GAIT associated proteins (L13a, EPRS, GAPDH) which together form a complex that bind GAIT motifs and prevent ribosomal translation of these mRNA (Figure 4.6).¹⁵⁰ In the context of IFN- γ release this response is considered a 'brake' or negative feedback mechanism for the acute inflammatory response.

This result was unexpected because the patients in the ‘red’ MOSAIC cluster (3) had significantly higher concentrations levels of IFN- γ than the other clusters (Figure 3.9 and 4.7). ‘Red’ cluster samples had high concentrations acute inflammatory response cytokines (TNF- α , IL-6, IL-8), which suggested that this IFN- γ -mediated negative feedback mechanism was failing in these patients.

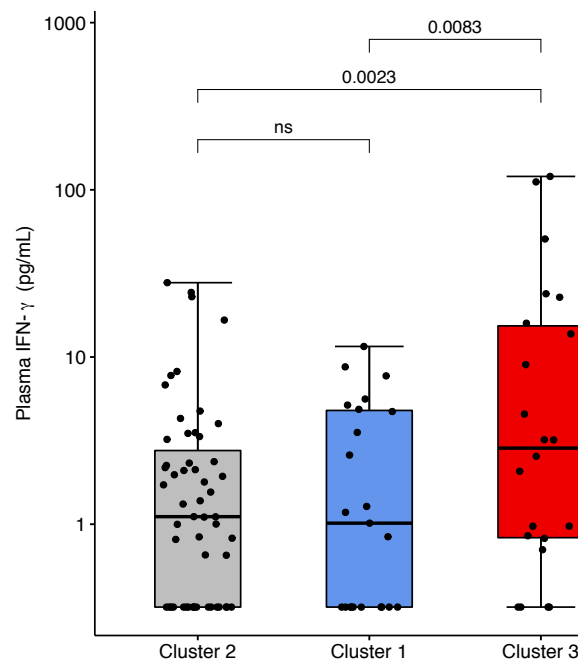


Fig. 4.7 Boxplot showing relative plasma levels of IFN- γ in each cluster from the MOSAIC study. The Kruskal-Wallis and post hoc Dunn's test were used to compare the differences between clusters. Samples from the ‘red’ cluster (3) had significantly higher IFN- γ levels than samples from both of the other clusters.

The role of IFN- γ in the context of influenza is complex. IFN- γ is released by CD8+ lymphocytes and NK cells in response to stimulation by IL-15. IL-15 was found to be raised in the patients with the 'red' MOSAIC cluster. Both NK and CD8+ T cells contribute to the lung injury seen in murine models of H1N1 infection and this pathology can be mitigated by stimulation with type 1 interferons.¹⁵¹

IFN- γ stimulation of cells via its receptor is also associated with activation of type 1 interferon anti-viral mechanisms. This signal is transduced via the JAK-STAT pathway which leads to activation of cyclin dependant kinase 5 (CDK5). CDK5 phosphorylates glutamyl-prolyl-tRNA synthetase (EPRS) releasing it from the multi-synthetase complex (MSC) and allowing it to initiate the GAIT complex.¹⁵² In addition, free EPRS activates the mitochondrial anti-viral system (MAVS) which leads to transcription of type 1 interferon related genes.¹⁵³ This entire pathway is considered to be part of an anti-RNA viral immune mechanism (Figure 4.8).

These mechanisms are of particular relevance here because the non-structural 1 (NS-1) protein of influenza A virus degrades STAT signalling, preventing the activation of CDK5 and so inhibiting release of EPRS from MSC.¹⁵⁴ This function of influenza A NS-1 protein serves as a viral immune evasion mechanism preventing cells from responding to IFN- γ by producing type 1 interferons. The implication here is that this function may precipitate an unchecked acute phase response by preventing the regulatory GAIT mechanism from acting.

In the microarray version used in the MOSAIC study there was no probe corresponding to the *EPRS* transcript, measurement of which might have provided additional verification of the role of this pathway in patients with severe influenza.

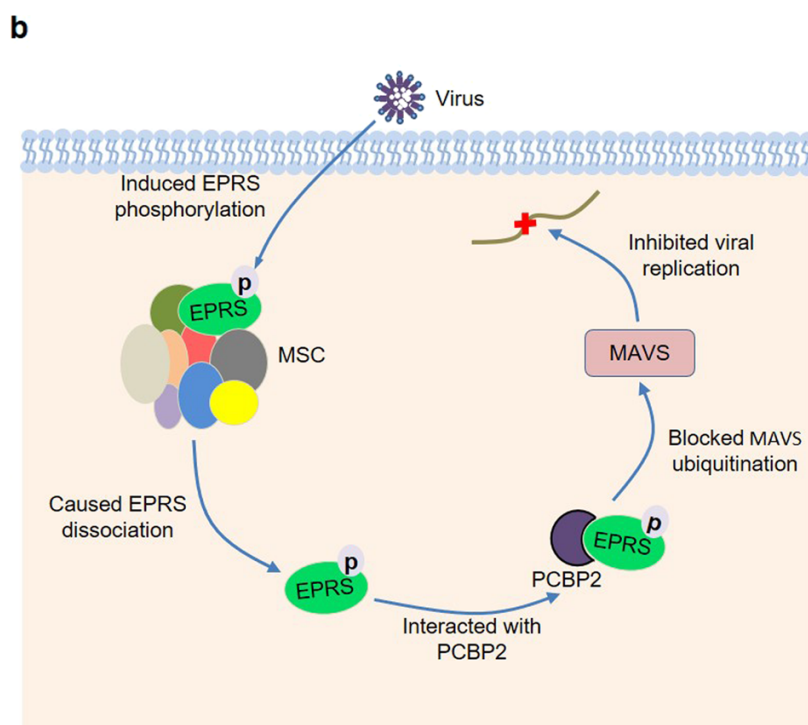


Fig. 4.8 Schematic representation of the role of EPRS in anti-viral immune mechanisms. Phosphorylation of EPRS at Ser990 releases it from MSC. EPRS with PCBP2 and blocks PCBP2-mediated MAVS ubiquitination. MAVS stimulates transcription of type 1 interferons.¹⁵⁵ Although the figure above states that viruses activate this mechanism, CDK-5, under the influence of IFN- γ , can also phosphorylate EPRS to activate this pathway.¹⁵²

Figure 3b taken from Anzheng Nie, Bao Sun, Zhihui Fu and Dongsheng Yu.

Roles of aminoacyl-tRNA synthetases in immune regulation and immune diseases

Cell Death Disease 10, 901 (2019).

<https://doi.org/10.1038/s41419-019-2145-5>

This figure is cited under under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
<https://creativecommons.org/licenses/by/4.0/>.

4.3 Linear discriminant analysis of integrated biological data

4.3.1 Combining module eigengene and protein biomarker values preserves the properties of clusters

Module eigengenes represent the first principal component of a gene module. The WGCNA library calculates the explained variance from each sample with respect to the first principal component of each gene module. This mathematical property, therefore, links every sample to each gene module. This is a powerful method of down-sampling the expression levels of several, or thousands, of highly correlated genes to a single value that can be managed like any other feature of a sample.

The module eigengene values were centred and scaled and then combined with the log-transformed, scaled protein biomarker values. The new data was projected into the principal component space and a biplot showed that the scaled eigengene values were orthogonal to the protein biomarker values. Figure 4.9 shows that the arrows labelled ‘ME’ are perpendicular to the arrows associated with cytokine labels. This was important because it was safe to presume that new data from module eigengenes was complementary and did not disrupt the variance of the protein biomarker data.

A three-dimensional plot of the first three principal components of this data showed that the relative locations of points in each cluster were relatively well preserved (Figure 4.10).

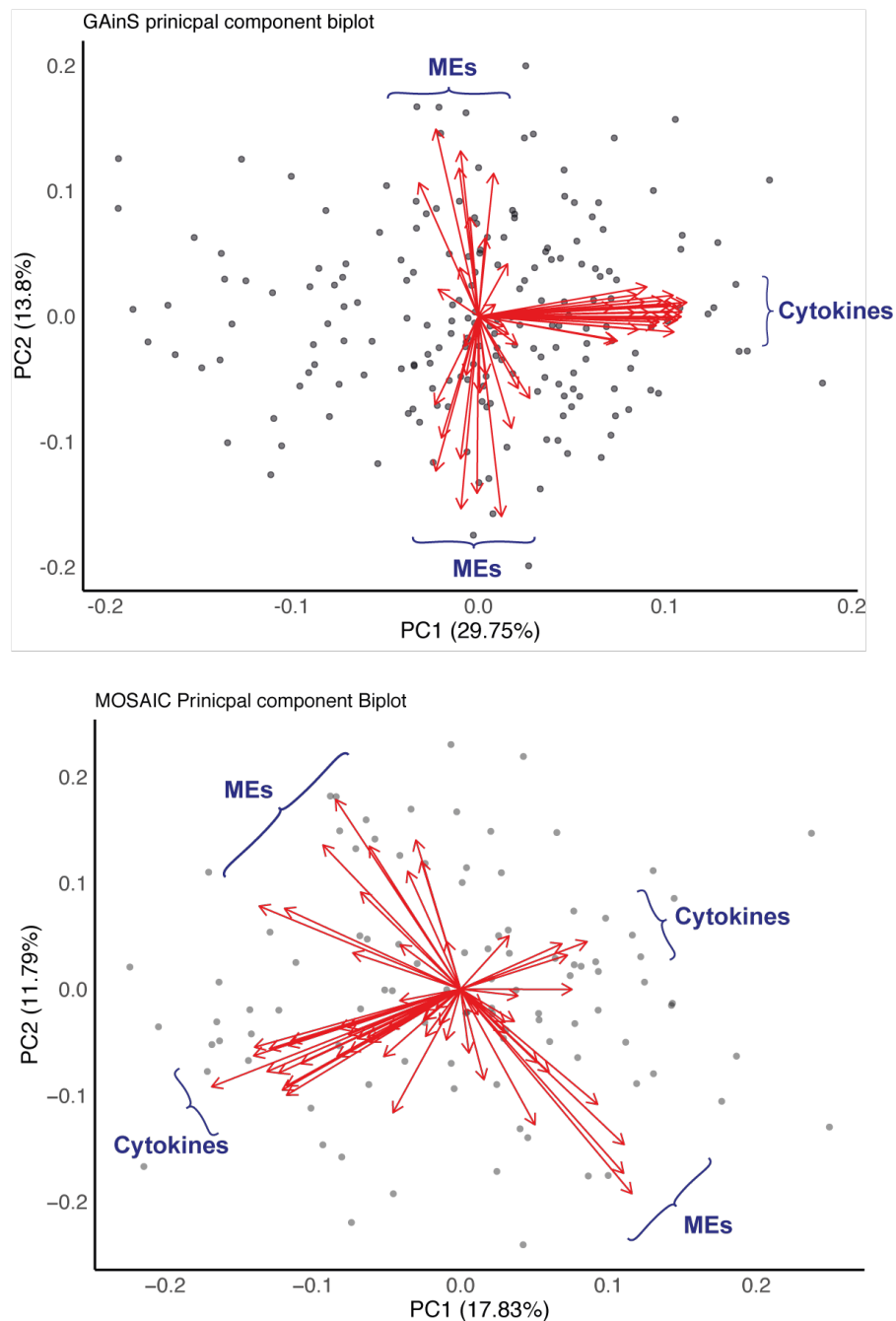


Fig. 4.9 Principal component plots with loadings (biplots) of the combined cytokine and eigengene values in both the GAINs and MOSAIC studies. The grey circles are individual samples projected into the PC space. The red arrows show the relative loadings of each contributory variable. Individual arrow labels have been omitted for clarity but can be seen in the Appendix E. In both plots the arrows labelled as cytokines are pointing in perpendicular directions to the module eigengene arrows (labelled 'ME'). This implies that variance of the data attributed to the eigengene values did not affect the variance of the protein biomarker data to large extent. The variance of data is important for statistical testing.

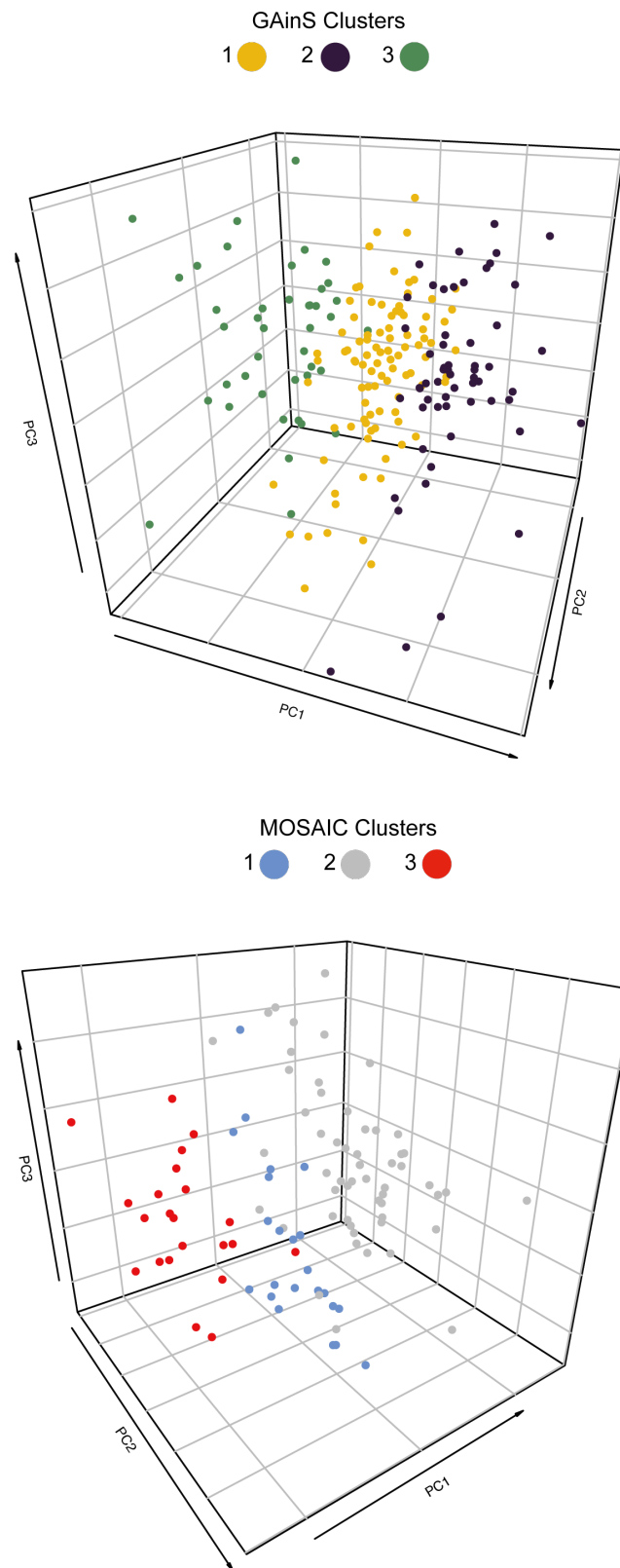


Fig. 4.10 Three dimensional projection of the first three principal components of the combined protein biomarker and eigengene values for both the GAINs and MOSAIC data. Each point is a patient sample and the colours represent the clusters determined by Ward linkage clustering of the cytokine data. The segmentation of the data points is preserved despite the augmentation of each sample with eigengene values.

4.3.2 Linear discriminant analysis of ARDS samples between different clusters from the GAINs study

The heatmaps of the cytokine levels and PCA plots of the combined cytokine and MEs from the GAINs study showed that the ‘purple’ (cluster 2) and ‘green’ (cluster 3) samples were the most divergent. The heatmaps also suggested that these two groups had global dysregulation of cytokine release, but in opposing directions.

Linear discriminant models were fitted for the subset of patients who had ARDS in each of these clusters. Linear discriminant analysis (LDA) is similar to PCA with respect to the underlying mathematics. The key difference is that LDA maximises the variance between classified data points.

Determination of whether this linear discriminant analysis (LDA) was suitable for this classification problem was conducted in two ways:

1. Calculate model performance for binary discrimination between pairs of clusters.
2. Calculate model performance for multi-class discrimination of all three clusters simultaneously.

Method 1 had to be applied when fitting models to small sample sizes with subsets of the data. Model performance was assessed using bootstrapped resampling to ensure robust results.

Method 2 was applied when models were fitted to all three clusters as the sample sizes were large. Class sizes were balanced and so leave-one-out cross-validation was used to ensure adequate model performance.

There were 196 samples in the GAINs study with protein biomarker and eigengene values. Of these, 71 were from patients with ARDS. 120 samples belonged to either the ‘purple’ or ‘green’ clusters. In the ‘purple’ and ‘green’ clusters, there were 38 samples in total with ARDS. 28 of the ARDS samples were from the ‘purple’ cluster.

Due to the relatively small sample size, the suitability of LDA for this classification problem had to be determined. An LDA model was first fitted to distinguish samples assigned to either the ‘purple’ or ‘green’ groups regardless of ARDS status. Model fit was assessed using leave one out cross-validation. In order to ensure that these results were robust, a bootstrapping method was also used. Using a larger sample size, by including patients without ARDS, also allowed for the data to be split into training and hold-out testing sets with a 75% splitting ratio.

LDA models were then fitted and tested by using the hold-out data and average performance was calculated using bootstrapped resampling with 1000 resamples.

The mean accuracy for this LDA model was equal to 0.94 (95% CI =0.80-0.99). Area under the receiver operating curve characteristic statistic (AUROC) was equal to 0.95 (95% CI = 0.88-1). These results implied the LDA method was an accurate classifier for these data.

An additional attribute of LDA models is their use in multiple classification where there are more than two predicted classes. An LDA model fitted to distinguish three classes would produce two decision boundaries. Figure 2.10b in Chapter 2.5.1 demonstrated this concept. For patients with ARDS in the GAINs study, an LDA model was fitted for to predict all three clusters and the visual representation of three class model fit are shown in Figure 4.11. The three clusters were seen to be linearly separable when transformed using the linear discriminant coefficients. Fitting a three class LDA model using the module eigengene values alone, without the protein biomarker values, did not demonstrate a clear separation between clusters.

A 75% split of the data into training and testing samples was used with bootstrapped resampling (1000 iterations) to assess the performance of a three class LDA predictor model. The mean accuracy of the model was equal to 0.88 (95%CI 0.82-0.92), and multiple class AUROC was equal to 0.92.ⁱ

ⁱusing the *multiclass.roc* function from the *pROC* R package. This function did not produce confidence intervals for multiple class AUROC statistics.

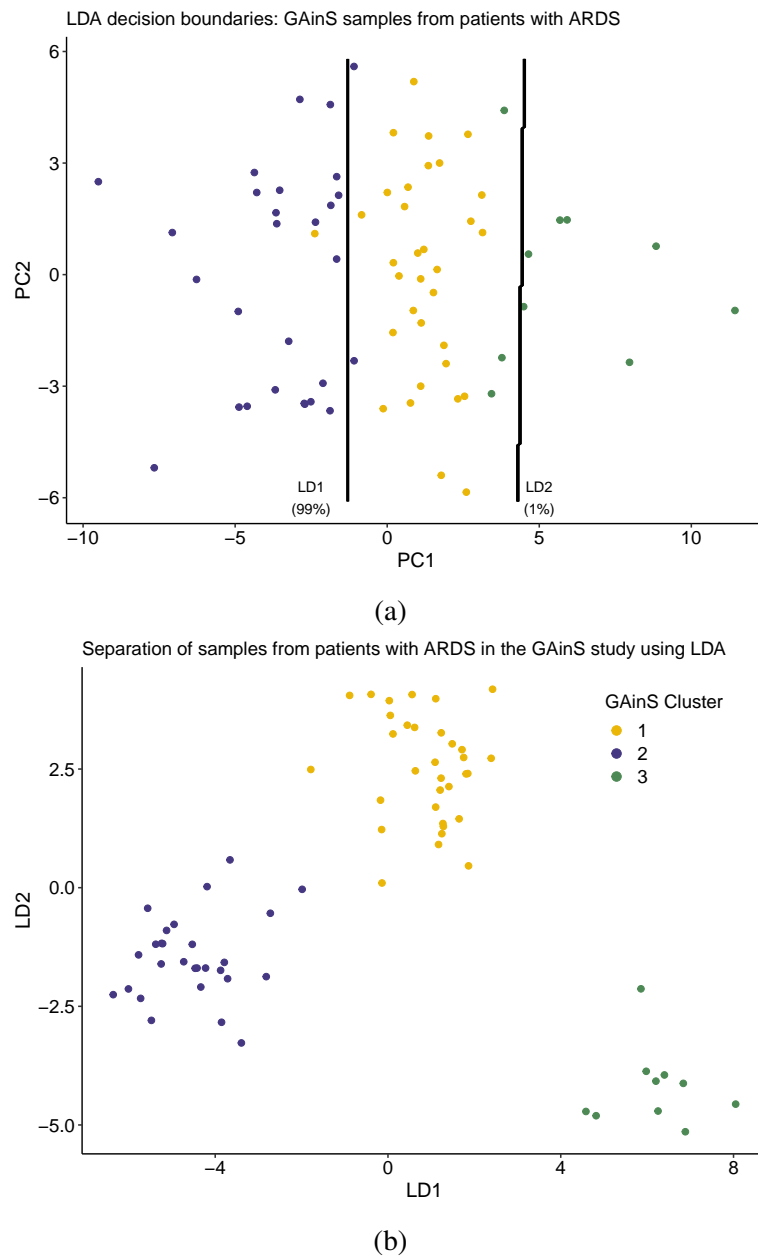


Fig. 4.11 LDA projections and decision boundaries for clusters found in ARDS patients recruited to the GAINs study. a.) shows an approximation of the decision boundaries that distinguish clusters if the data is projected into the PC1-PC2 subspace. The real boundaries are hyperplanes that cannot be easily visualised. b.) show the results of linear transformations $(\mathbf{X} \times \mathbf{L}_{D1})$ and $(\mathbf{X} \times \mathbf{L}_{D2})$ of the data points (\mathbf{X}) , using matrix multiplication, by the LD1 (\mathbf{L}_{D1}) and LD2 (\mathbf{L}_{D2}) coefficients. This linear transformation results in projection of these data points to new co-ordinates which are shown here. Each cluster grouping is well separated. (b.) is a more accurate representation of the LDA process compared with the approximations shown in (a.).

A more straightforward classification approach using logistic regression was unsuccessful; either the function did not converge on stable solutions or there were no predictions in the minority class. In contrast, the LDA models consistently predicted minority class instances. Poor minority class prediction is a frequently encountered problem with classification models.

Improving model performance was not the objective of this analysis, which was to determine the most important mechanisms that were defining ARDS in each of these clusters. In this context, this method was adequate to address the question of which variables were the most important discriminators between subtypes of ARDS.

4.3.3 Neutrophil activation is an important discriminator of ‘purple’ and ‘green’ GAIN clusters in ARDS samples

The same methods as above were used to fit and test an LDA model to classify membership of the ‘purple’ and ‘green’ clusters in samples from ARDS patients (38 samples in total). Accuracy for this fitted model was equal to 0.92 (95% CI: 0.62-1) and AUROC was equal to 0.94 (95% CI 0.84-1), which implied that this model performed well, even with small sample sizes.

After establishing adequate model performance, the key discriminating features between ARDS samples in the ‘purple’ and ‘green’ clusters were calculated. For this process, a full model, with no split of the data for testing, was used. The discriminant coefficients from this full model were then extracted and ranked. The coefficients of the linear discriminator were the linear combinations of the contributing variables that described the decision boundary between clusters. All the variables in the model contributed to this decision boundary but each variable’s relative contribution was described by the magnitude of its linear discriminant coefficient.

The ten highest discriminators, ranked by magnitude are shown in Figure 4.12. A plot showing all of the ranked variables is shown in Appendix Figure F.1. CXCL9 (MIG) and TNFR-2 were the highest ranking discriminant protein biomarkers. The highest ranking gene modules were the ‘royal blue’ and ‘black’ modules containing 104 and 321 labelled transcripts respectively. The ‘royal blue’ module did not enrich for any particular biological pathway. Nor did the ‘dark orange’ module which had a coefficient of similar magnitude to the ‘royal blue’ module. Enrichment of a combined list of transcripts from both of these modules together did not identify a significant biological pathway either. Interestingly, the ‘dark orange’ module was adjacent to the ‘black’ module on the gene module dendrogram

(Figure 3.18) suggesting these two modules may share related genes. The ‘black’ module, consisting to 321 transcripts, enriched for the biological process ‘neutrophil activation involved in immune response’ (GO:0002283, adjusted $p = 7.9 \times 10^{-6}$).

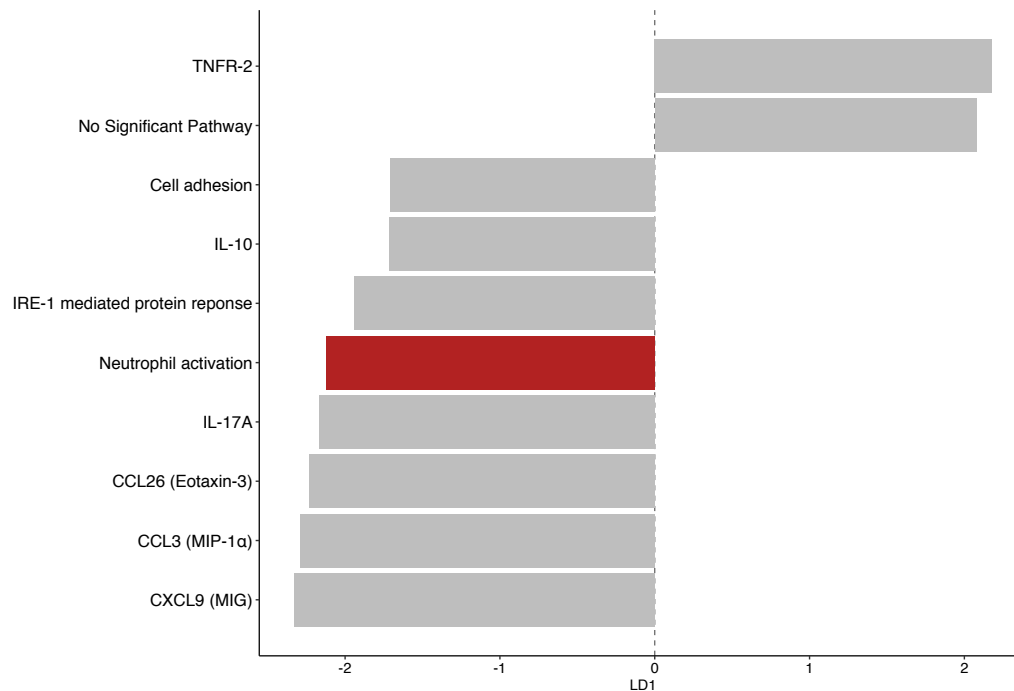


Fig. 4.12 Top ten ranked discriminators for ARDS samples from the GAINs study between ‘purple’ and ‘green’ clusters. These discriminators were selected by their relative magnitude. LDA is insensitive to directionality of the classifier labels. For example if the ‘purple’ and ‘green’ cluster labels were reversed (0 to 1, from 1 to 0) the discriminators still rank in the same order with the same magnitude. The gene module labelled ‘no significant pathway’ was coloured ‘royal blue’ and contained 104 labelled transcripts. The top ranking gene modules that was associated with a biological process is highlighted in red (‘neutrophil activation’). A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure F.1

In view of the large difference in protein biomarker concentrations between these two clusters, enrichment for genes related to neutrophil activation was anticipated. This finding provided some independent verification that the processes driving dysregulated cytokine release in these patients involved neutrophil activation. Although, in the context of sepsis, neutrophil activation was not an unexpected finding.

The ‘dark turquoise’ module which was ranked just below the black module, contained 89 transcripts and enriched for ‘IRE-1 mediated protein response’ (GO:0036498, adjusted $p = 0.008$). Inositol-requiring enzyme 1α (IRE-1) is an endoribonuclease that is released in

response to endoplasmic reticulum (ER) stress as part of the ‘unfolded protein response’. Once activated it acts as a transcription factor that facilitates production of chaperone proteins to ease ER stress.¹⁵⁶ Its role here probably reflects cellular stress or failure of protein homeostasis.

4.3.4 Ranked discriminators of the ‘yellow’ and ‘green’ GAIN clusters in ARDS samples involve transcripts with important roles in immune function

To determine the discriminators between the ‘yellow’ and ‘green’ clusters from the GAIN study, the same methods as sections 4.3.2 and 4.3.3 were used to fit an LDA model. The model accuracy was equal to 0.86 (95% CI 0.57 - 0.98) and AUROC was equal to 0.79 (95% CI: 0.45-1). The total sample size was 43.

The ten discriminators with the largest coefficients, determined by fitting a model on the full, non-split data, are shown in Figure 4.13. The highest ranked protein biomarkers were IL-17a and IL-1 β . The highest ranked gene module was the ‘light green’ module which consisted of 105 transcripts and enriched for the process ‘regulation of protein phosphorylation’ (GO:0001934, adjusted $p = 0.002$).

Although this ontology label may have suggested little relevance to ARDS or sepsis, there were only five transcripts which determined enrichment to this pathway:

- *CSF1R*
- *CD74*
- *CD4*
- *ARRB1*
- *ENG*

All of the proteins that are encoded by these transcripts have important immune-related functions.

CSF1R is a receptor for colony stimulating factor (CSF1) which is a cytokine that is important for macrophage differentiation.¹⁵⁷ CD74 protein is a chaperone for the human leucocyte antigen (HLA) class II histocompatibility protein. It therefore plays an important role in antigen presentation but has additional roles in immunology and physiology. It acts a receptor for macrophage inhibitory factor (MIF) and directly influences angiotensin II type 1

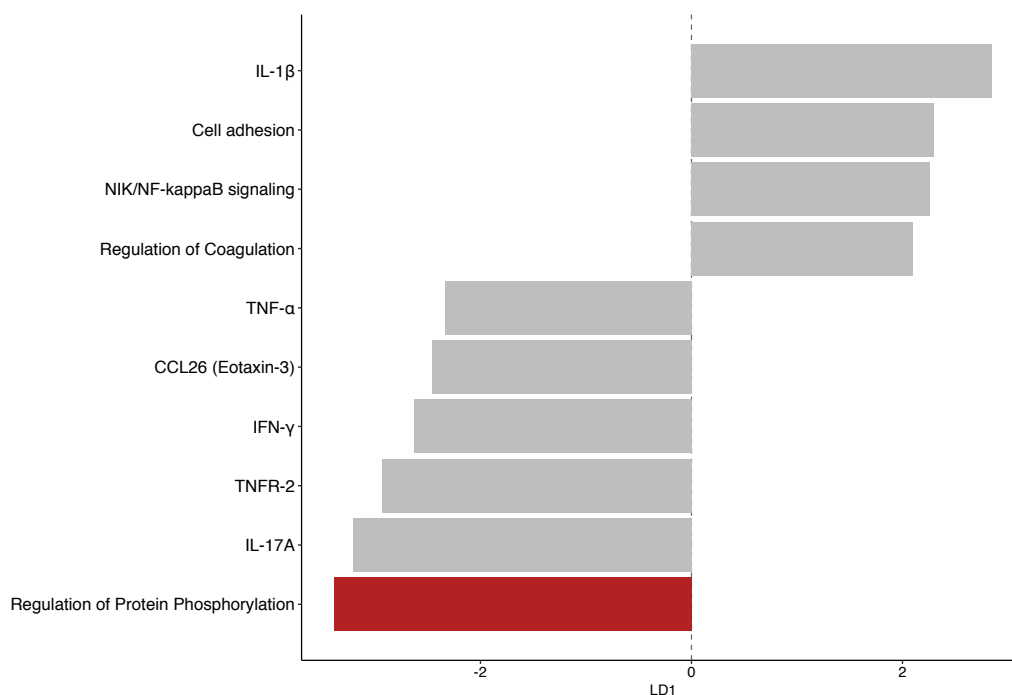


Fig. 4.13 Top ten ranked discriminators for ARDS samples from the GAINs study between ‘yellow’ and ‘green’ clusters. These discriminators were selected by their relative magnitude. The top ranking gene module that was associated with a biological process is highlighted in red (‘regulation of protein phosphorylation’), which contained transcripts that play important roles in immunity.

A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure F.2

receptors (AT1) by inhibiting trafficking to the cell surface, thus promoting degradation.¹⁵⁸ *ARRB1* codes for the protein β -arrestin 1 which is important in the regulation of activated g-protein coupled receptors and is ubiquitously expressed. It plays an important role in cell survival and immunological signalling pathways.¹⁵⁹ Endoglin (ENG) is a glycoprotein involved in vascular endothelial integrity. Endothelial integrity is often lost in sepsis resulting in endothelial leak and circulating hypovolaemia. Mutations in this gene are associated with the disease hereditary haemorrhagic telangiectasia (HHT) which is associated with recurrent micro and macro-haemorrhage due to vascular dysplasia.¹⁶⁰ Circulating endoglin concentrations are higher in patients with septic shock than healthy controls and this is thought to be due to shedding mediated by matrix metalloproteinases released from macrophages.¹⁶¹

4.3.5 Ranked discriminators of the ‘yellow’ and ‘purple’ GAINs clusters in ARDS samples do not identify a plausible mechanism to account for their differences

The number of samples with ARDS in the ‘purple’ and ‘yellow’ clusters was 61. An LDA model was fitted to discriminate these two clusters in the same manner as before. Model performance was worse compared with the models fitted between the other clusters previously, despite the relatively larger sample size. The mean accuracy was equal to 0.67 (95% CI: 0.38 - 0.88) and AUROC was equal to 0.66 (95% CI: 0.41-0.92).

A possible explanation might be the relative adjacency and overlap of these two clusters. If the ‘yellow’ and ‘purple’ points in Figure 4.11 are projected solely onto the LD1 axis, one ‘yellow’ cluster sample would overlap a ‘purple’ cluster sample. From the LD2 perspective, there may be as four overlapping samples between clusters. These overlapping samples would introduce instability into the model, especially if they were repeatedly resampled. Model instability might also explain why the discriminator coefficient values for this model, between ‘purple’ and ‘yellow’ clusters, are of greater magnitude and range (-22 to 15) compared with the ranges of the coefficients from the other fitted models (-3 to 3). This can be observed by comparing the values of the horizontal axes in Figures 4.14, 4.12 and 4.13.

The ten discriminators with the largest coefficients, determined by fitting a model on the full, non-split data, are shown in Figure 4.14. The highest ranked cytokines were CXCL11 (I-TAC) and CXCL9 (MIG). The highest ranked gene module was the ‘turquoise’ module which consisted of 2,820 transcripts and enriched for the process ‘viral mRNA translation’ (R-HSA-192823, adjusted $p = 1.9 \times 10^{-41}$). The ‘turquoise’ module transcripts enriching for this process contained many ribosomal RNA coding genes. The high ranking ontology labels for this ‘turquoise’ module all contained a common set of 13 transcripts (*RPL4*, *RPL30*, *RPL3*, *RPL32*, *RPL31*, *RPL34*, *RPLP1*, *RPLP0*, *RPL8*, *RPL10A*, *RPL9*, *RPL6*, *RPL7*). The ontologies with this set of 13 transcripts had general themes relating to mRNA translation. Inferring viral infection from this module label of ‘viral mRNA translation’ would therefore be questionable.

Of note was the presence of the ‘black’ module in this list of highest ranked variables. The same module was important in the ‘purple’ and ‘green’ clusters (Figure 4.12) and enriched for the term ‘neutrophil activation’ (GO:0002283, adjusted $p = 9.8 \times 10^{-6}$).

In view of the instability of the predictor model the significance of these identified pathways were uncertain although they might suggest a different inflammatory profile in the patients that belonged to the ‘yellow’ cluster. There may have been a contribution of viral infection to the gene expression profile in these samples but this would require verification with serology or other biomarkers.

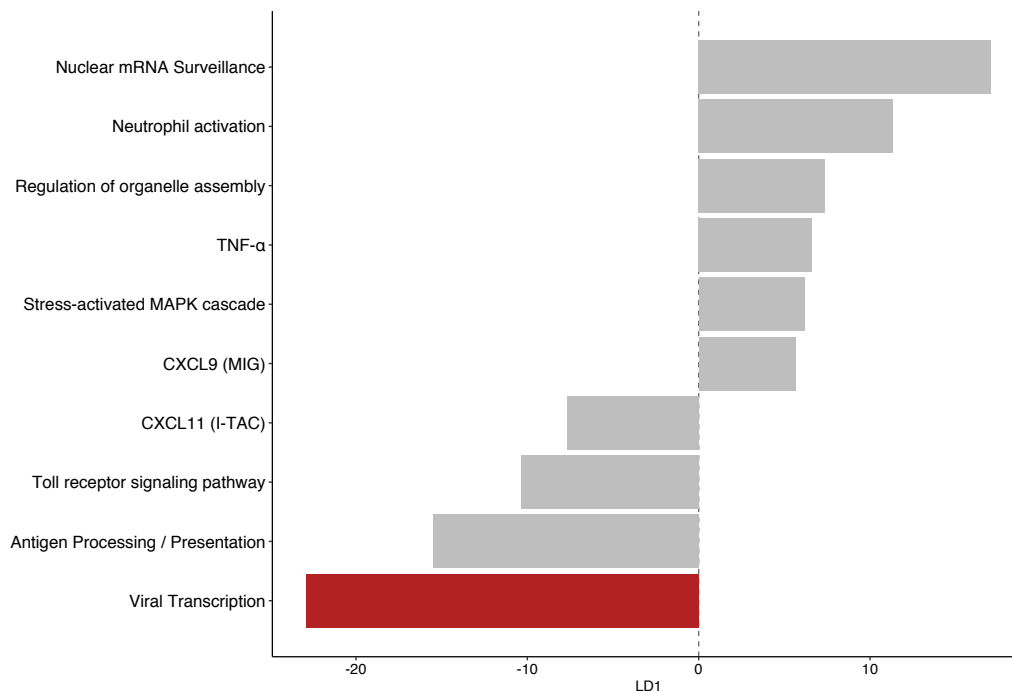


Fig. 4.14 Top ten ranked discriminators for ARDS samples from the GAINs study between ‘purple’ and ‘yellow’ clusters. These discriminators were selected by their relative magnitude. The top ranking gene module that was associated with a biological process is highlighted in red (‘viral mRNA translation’). Of note was the recurrence of the ‘black’ gene module which enriched for ‘neutrophil activation’ as this same module was discriminant between the ‘purple’ and ‘green’ clusters.

A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure F.3.

4.3.6 Further enrichment of key modules that discriminated the GAINs ‘purple’ and ‘green’ clusters identifies important sub-networks

The ‘black’ module which enriched for neutrophil activation was the most discriminant gene module between the ‘purple’ and ‘green’ samples from patients ARDS in the GAINs study (Figure 4.12). The eleventh highest ranking discriminator from this model was the ‘dark orange’ gene module which did not enrich for any given pathway or ontology at statistically significant level. This module was however adjacent to the ‘black’ gene module in the dendrogram of modules distances (Figure 3.18). In view of their high rank and adjacency the transcripts from these two modules were combined and submitted for further enrichment using a proprietary tool that was temporarily available to the project called Metabase (Clarivate Analytics Limited).

This tool queried a network of curated resources for known literature-based interactions between submitted genes. The threshold for this function was set to two publications so that at least two independent references had to cite an interaction between genes for these to be considered valid. Gene labels that satisfied this minimum screening and then checked for additional interactions with the other genes in the submitted list. Groups of more than two genes are referred to as sub-networks.

Sub-networks were pruned using an edge “betweenness” measure described in the Girvan-Newman algorithm.¹⁶² Edge betweenness refers to identification of nodes that are the least central in a community and the paths between them and other local communities in a network. If communities are only connected to an adjacent community by a few short paths then this path is referred to as having high edge betweenness. Pruning this edge will isolate the two communities, leaving sub-networks. The splitting and merging of networks allows for calculation of modularity based on the membership of the original network, prior to pruning. Nodes with multiple edges linking them will be emphasised by this process.

Most of the identified interactions after pruning were simple pair-wise gene interactions. To identify the large communities of genes, four member was set as a threshold for ongoing analysis. Nine sub-networks were revealed then this tool and these processes were applied to the transcripts in the ‘black’ and ‘dark orange’ modules from the GAINs transcriptomic data (Figure 4.15).

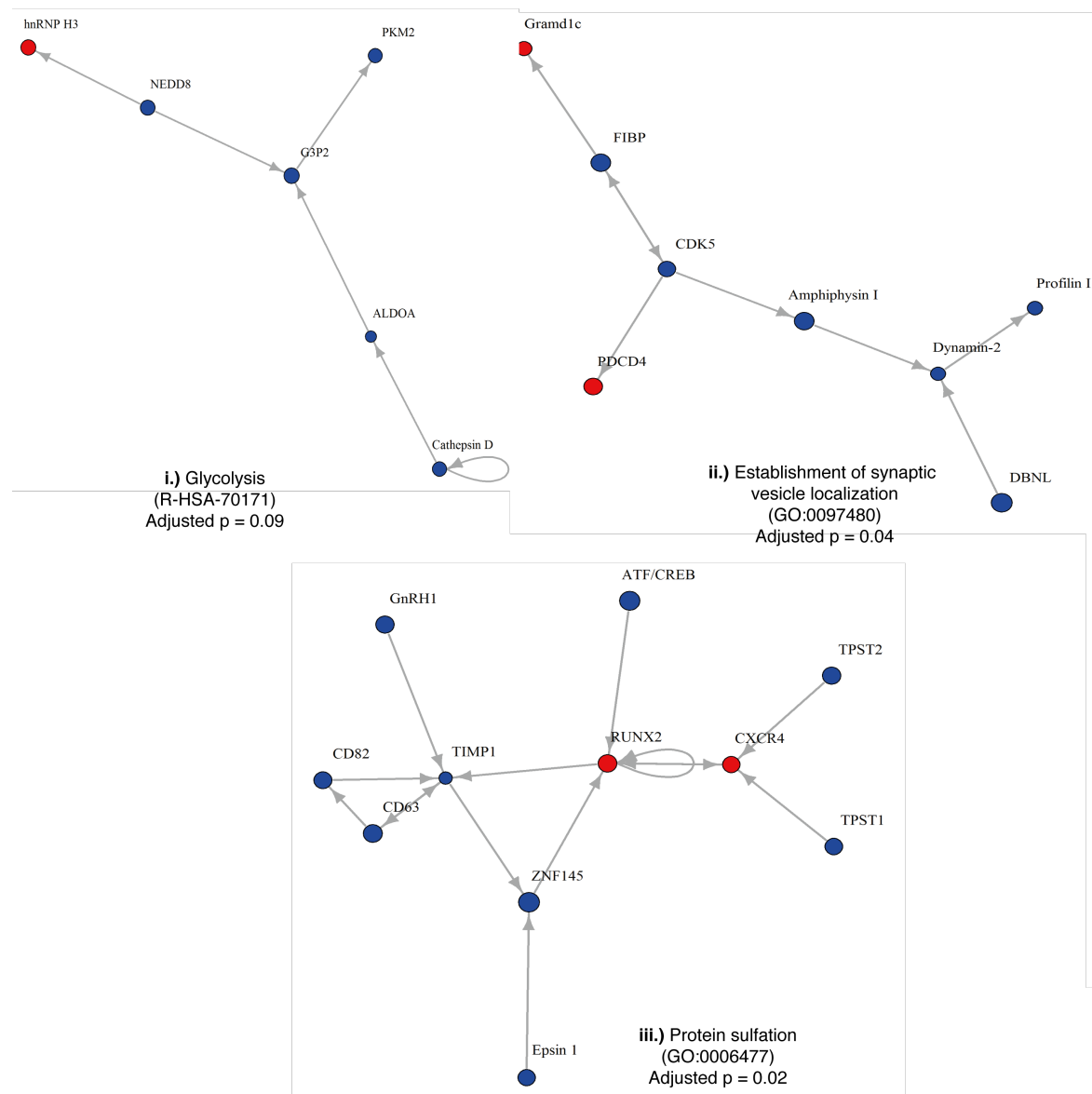
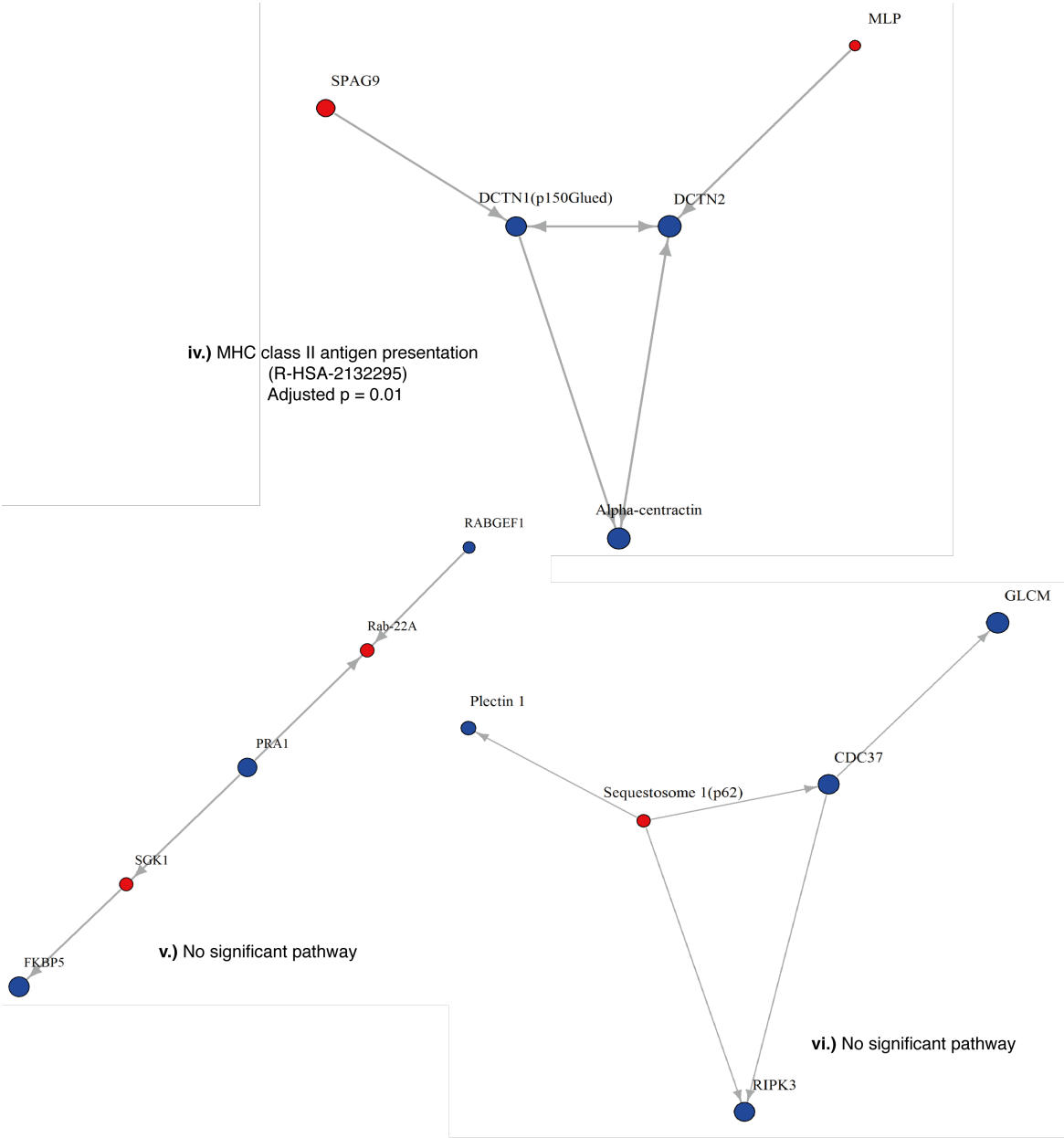
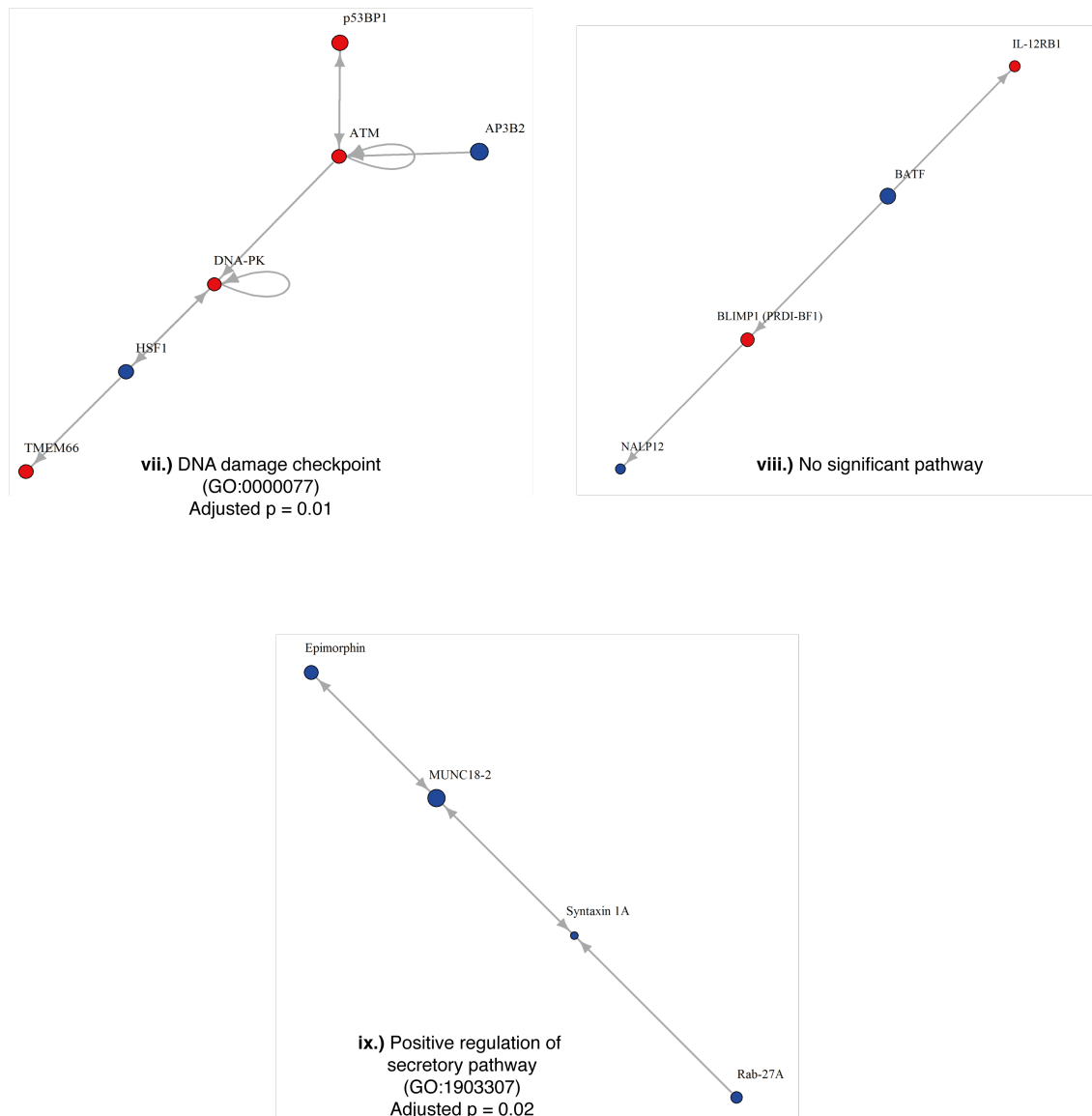


Fig. 4.15 (a) Metabase-identified sub-networks.
Full caption with sub-figure (c)



4.15 (b) Metabase-identified sub-networks.
Full caption on following page.



4.15 (c) Sub-networks identified by Metabase (Clarivate Analytics Limited) from the transcripts contained in the ‘black’ and ‘dark orange’ gene modules from the GAINs study. The tool used known annotations of interactions between genes and proteins in the submitted list to build a network which was pruned to isolated sub-networks. Sub-networks of four or more genes are shown here. Of particular note is (ix) which enriched for the pathway positive regulation of regulated secretory pathway. This sub-network contained genes associated the familial hemophagocytic lymphohistiocytosis (HLH) which is associated with severe immune dysfunction and cytokine release which may result in multi-organ failure. It often presents as a clinical syndrome which resembles severe septic shock.

Of particular interest are sub-networks (v), (vi) and (ix). Sub-network (ix) enriched for the process ‘positive regulation of regulated secretory pathway’ (GO:1903307, adjusted $p = 0.02$). Two genes in this pathway (*MUNC18-2*, *Rab27a*) are associated with the familial forms of haemophagocytic lymphohistiocytosis (HLH). HLH is severe form of immune dysfunction characterised by excessive macrophage activation, dysregulated cytokine release, haematophagocytes in the bone marrow and multi-organ failure. Patients have pancytopenia, high levels of serum ferritin, triglycerides and may develop a profound coagulopathy.¹⁶³ If patients require management in intensive care the reported hospital mortality ranges between 52% and 68%.¹⁶⁴ This gene module was associated with the ‘purple’ cluster characterised by globally raised cytokines. The association of this cluster with genes that may play a role in dysregulated cytokine release in the context of sepsis is a possible mechanism that would account for the protein biomarker profile observed in these patients.

HLH can be primary or secondary. Primary HLH occurs in children and is related to inherited mutations. Secondary HLH occurs in adults and is usually caused by malignancy (lymphoma), infection (EBV, CMV, HIV), connective tissue disorders or is idiopathic. HLH is sometimes considered a ‘sepsis-mimic’ as it shares clinical features which may resemble sepsis: fever, multi-organ dysfunction, pancytopenia.

The pathogenesis of HLH relates to dysregulated NK cell and cytotoxic CD8+ T-cells, which cause proliferation of lymphocytes and histiocytes (tissue macrophages). The gold-standard test is the identification of haematophagocytes on bone marrow aspiration. Haematophagocytosis refers to macrophages engulfing erythrocytes, but frequently lymphocytes, platelets and other precursor cells are seen to be engulfed. The diagnostic criteria for HLH, described in 2004 by the Histiocyte Society, are shown in Table 4.1.¹⁶⁵ The identification of genes related to HLH in the context of severe cytokine dysregulation in patients with sepsis is independent validation of this cytokine profile in these patients and suggest an explanatory mechanism.

Sub-network (v) did not enrich for a pathway or process, however it contained two important genes related to glucocorticoids (steroids). FKBP5 (FK506 binding protein 5) and SGK1 (serum and glucocorticoid-regulated kinase 1) are both important in glucocorticoid-related biological pathways. FKBP5 modulates glucocorticoid-receptor activity, with higher levels of this protein associated with reduced glucocorticoid transcription. FKBP5 transcription is increased following activation of glucocorticoid receptors.¹⁶⁶ SGK1 has a number of intracellular roles. The transcription of this gene is under the control of glucocorticoids. Glucocorticoids are frequently administered to patients with septic shock. If FKBP5 was up-regulated in these patients, this might have been in response to glucocorticoid therapy.

Sub-network (vi) contained RIPK3, an important protein in necroptosis. Necroptosis is a form of programmed cellular necrosis that is different to apoptosis and passive necrosis. Necroptosis may be triggered by death receptors in response to damage associated molecular patterns (DAMPs) or interferons. It is increasingly recognised as playing a key role in inflammatory disorders and the response to viral and bacterial infection. Activation of RIPK3 is the key step in necroptosis and so its presence in the context of sepsis and ARDS is of interest.¹⁶⁷

Parameter
1. Fever
2. Splenomegaly
3. Cytopenia of at least two cell lines
- Hb < 90 g/L
- Plts < 100×10^9 /L
- Neutrophils < 1×10^9 /L
4. Raised triglycerides and/or low fibrinogen
- fasting triglycerides > 3.0 mmol/L
- fibrinogen < 1.5 g/L
5. Haematophagocytosis, demonstrated
in bone marrow, spleen or lymph nodes
6. Low or absent NK cell activity
7. Ferritin > 500 μ g/L
8. Soluble CD25 (IL-2 receptor) > 2,400 iu/mL

Table 4.1 HLH diagnostic criteria as described by the Histiocyte Society (HLH-2004 protocol). Five of the eight above criteria are required for a diagnosis of HLH.

4.3.7 Linear discriminant analysis of samples between different clusters from the MOSAIC study

The same approach as Section 4.3.2 was used for the MOSAIC data. Here there was no specific label for ARDS, so severe respiratory failure (respiratory SOFA ≥ 3) was used to ensure the identified mechanisms were applicable to the most unwell patients. In contrast to the GAINs study, the MOSAIC sample size for combined immune mediator and transcriptomic results was smaller (100 samples). The clusters were also smaller and unbalanced (grey: 52, blue: 23, red: 22). For the subset of samples with respiratory SOFA ≥ 3 the number in each group were grey: 21, blue: 21, red: 20.

A multi-class LDA model, fitted with leave-one-out cross-validation, had an accuracy of predicting membership of each of these classes equal to 0.80 (95% CI: 0.76-0.84), and multi-class AUROC equal to 0.83. The projection of this model into the PCA 1/2 subspace on to the LDA dimensions is shown in Figure 4.16. It can be seen from this figure that projection using the first linear discriminant separates all three cluster effectively. The performance of the multi-class LDA model might therefore have been expected to better than the results seen here. This is probably due to the small number of samples (approximately 20 in each cluster). For a multi-class classification model, where the expected probability of success for a given prediction is the prior probability of the largest class (0.34), this performance was considered to be satisfactory.

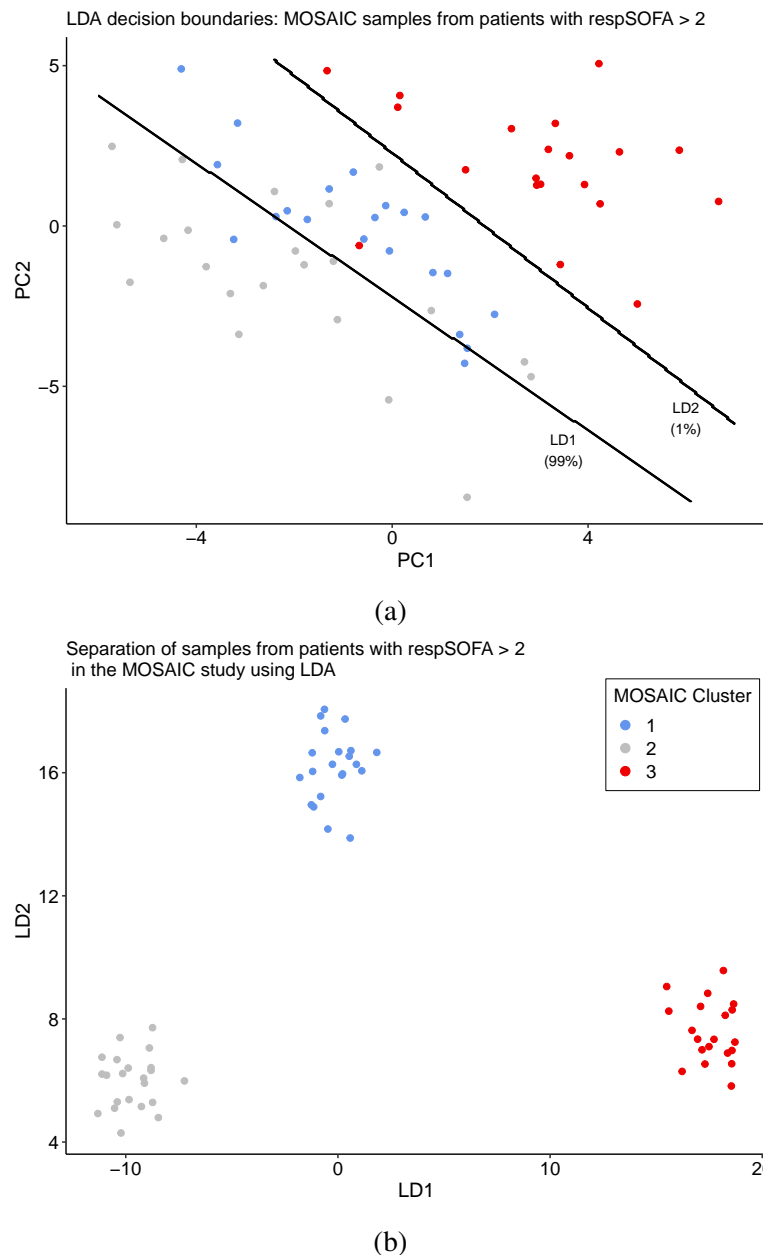


Fig. 4.16 LDA projections and decision boundaries for clusters in samples from patients with respiratory SOFA > 2 recruited to the MOSAIC study. a.) shows an approximation of the decision boundaries that distinguish clusters if the data is projected into the PC1-PC2 subspace. The real boundaries are hyperplanes that cannot be easily visualised. b.) show the results of linear transformations ($\mathbf{X} \times \mathbf{L}_{D1}$) and ($\mathbf{X} \times \mathbf{L}_{D2}$) of the data points (\mathbf{X}), using matrix multiplication, by the LD1 (\mathbf{L}_{D1}) and LD2 (\mathbf{L}_{D2}) coefficients. This linear transformation results in projection of these data points to new co-ordinates which are shown here. Each cluster grouping is well separated with respect to the first linear discriminant axis (LD1). (b.) is a more accurate representation of the LDA process compared with the approximations shown in (a.).

4.3.8 Neutrophil degranulation discriminates the ‘grey’ and ‘red’ MO-SAIC clusters

The same methods that were used to differentiate clusters found in the GAINs study were used here for the clusters identified in the MOSAIC study: leave one out cross validation, and bootstrapped resampling to calculate confidence intervals.

An LDA model was fitted to differentiate ‘red’ and ‘grey’ MOSAIC clusters using samples from patients that had a respiratory SOFA score ≥ 3 . Mean accuracy for this fitted model was equal to 0.94 (95% CI: 0.99-0.88). AUROC was equal to 0.83. Given there were only approximately 20 samples in each of these clusters, performance of this model was reassuring.

The discriminant coefficients from the fitted model were ranked by magnitude and these results for the ten highest variable can be seen in Figure 4.17. A line plot showing the ranking of all variables is shown in Appendix Figure G.3. The most discriminant cytokine for this model was TNF- α . The highest ranked gene module was the ‘black’ module which enriched for the ontology ‘neutrophil degranulation’ (GO:0043312, adjusted $p = 2.5 \times 10^{-12}$). The ‘red’ cluster had much higher levels of TNF- α and other cytokines associated with neutrophil activation (Figure 3.9 and Figure 4.5) so this was considered a plausible finding.

The transcripts in the ‘black’ module also enriched for processes associated with glycolysis and carbohydrate metabolism. Inappropriate activation of metabolic pathways has previously been described in sepsis by the GAINs and MARS research groups.^{96,98} Other plausible explanations include: metabolic insufficiency due to inadequate substrate availability in critically unwell patients, and an epi-phenomenon of a large number of metabolically active neutrophils circulating in these patients. Neutrophils have few mitochondria and are recognised as predominantly glycolytic.¹⁶⁸

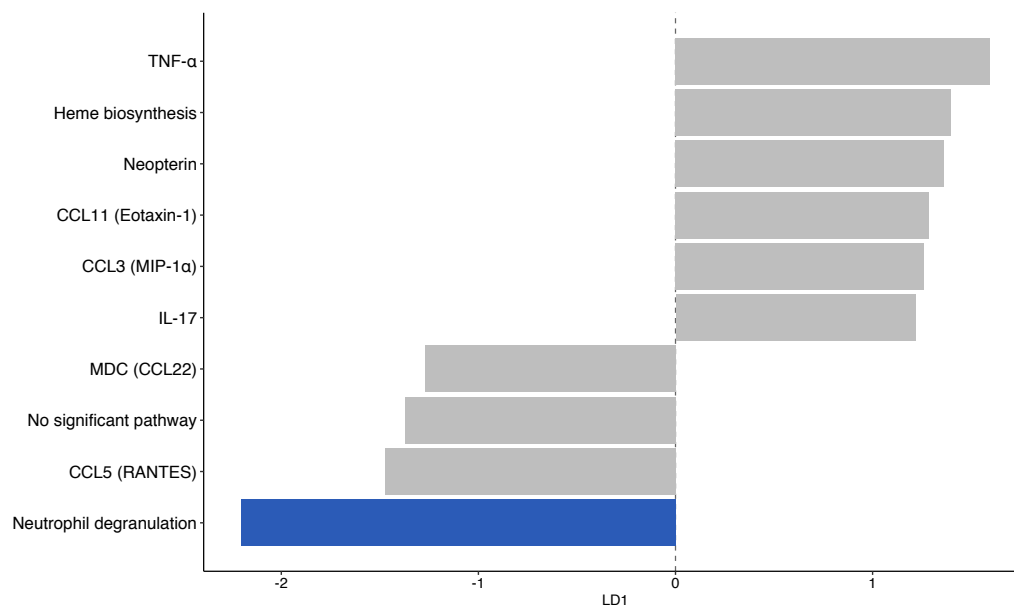


Fig. 4.17 Top ten ranked discriminators for samples with respiratory SOFA score > 2 from the MOSAIC study between ‘grey’ and ‘red’ clusters. These discriminators were selected by their relative magnitude. The top ranking gene module that was associated with a biological process is highlighted in blue (‘neutrophil degranulation’).

A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure G.3.

4.3.9 Regulation of SLITs and ROBOs discriminates the ‘red’ and ‘blue’ clusters from the MOSAIC study

An LDA model was fitted to differentiate the ‘blue’ and ‘red’ MOSAIC clusters using samples from patients with respiratory SOFA score greater than 2. Mean accuracy for this fitted model was equal to 0.93 (95% CI 0.99-0.88). AUROC was equal to 0.9. This result was consistent with the model being a robust classifier.

The discriminant coefficients were ranked by magnitude and these results for the ten highest ranking variables can be seen in Figure 4.18. The most discriminant cytokine for the model was TNF- α . The highest ranked gene module was the ‘royal blue’ module which enriched for the ontology ‘regulation of SLITs and ROBOs’ (R-HSA-9010553, adjusted $p = 6.3 \times 10^{-4}$). SLIT-ROBO signalling is considered important in neuronal axon development and growth. There are at least five human ROBO receptors.

Slit-2-ROBO-4 signalling has recently been implicated as playing an important role in pulmonary endothelial leak in mice with sepsis and influenza infection.¹⁶⁹ The same mechanism has also been shown to mediate endothelial integrity using *in vitro* models. One model used pulmonary derived vascular endothelial cells (PMVECs) infected with hantavirus and another emulated transfusion-associated lung injury (TRALI) by treating PMVECs with anti-human neutrophil antigen antibodies.^{170,171}

Slit-2 binding to ROBO-4 receptors on pulmonary vascular endothelial cells preserves surface VE-cadherin which maintains endothelial integrity inhibiting leak of inflammatory infiltrate into the alveolar space (Figure 4.19). The implication here is that this mechanism plays a role in influenza infection of humans that develop severe respiratory failure. This mechanism has not previously been described in human influenza infection.

The other highly ranked discriminators here were consistent with the immune mediator profiles of each cluster; the ‘red’ cluster had had high levels of TNF- α , procalcitonin and CXCL10. Another highly ranked module in this model was the ‘midnight blue’ module which enriched for ‘antimicrobial humoral response’ (GO:0019730, adjusted $p = 2.1 \times 10^{-9}$). There were 10 genes from the module that were associated with this ontology that had important roles in immunity that were strongly related to neutrophil activity (Figure 4.2). Of these ten transcripts, two (*LCN2* and *BPI*) have previously been identified by Kangelaris *et al* (2015) as being up-regulated in patients with ARDS. In their study, the authors confirmed that four of the fifteen genes that were up-regulated had increased expression using qPCR. *BPI* and *LCN2* were two of these four genes. This adds a degree of external validation to the results shown here.

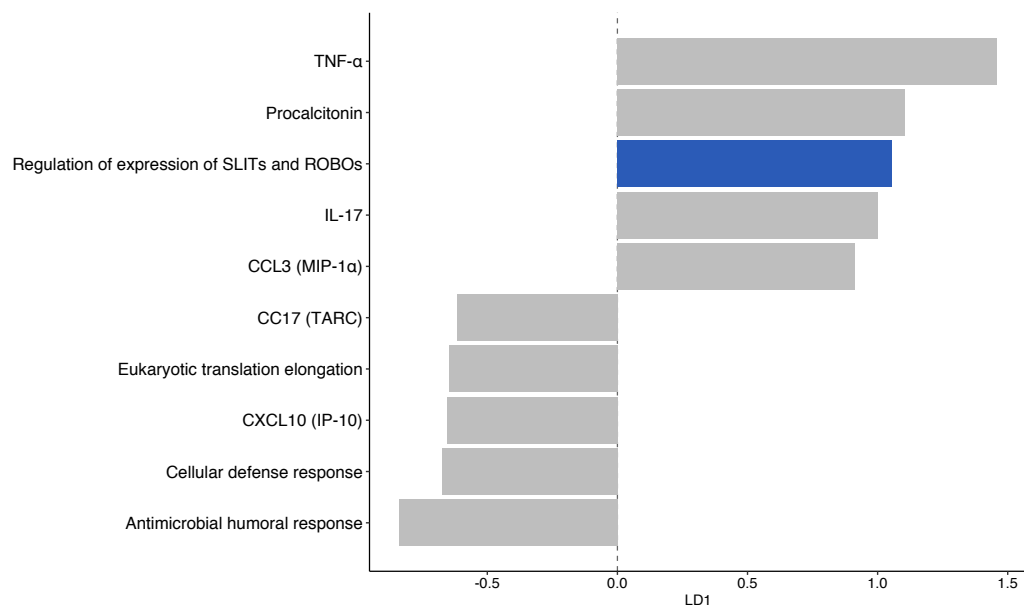


Fig. 4.18 Top ten ranked discriminators for samples with respiratory SOFA score > 2 from the MOSAIC study between ‘blue’ and ‘red’ clusters. These discriminators were selected by their relative magnitude. The top ranking gene module that was associated with a biological process is highlighted in blue (‘regulation of expression of SLITs and ROBOs’).

A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure G.2.

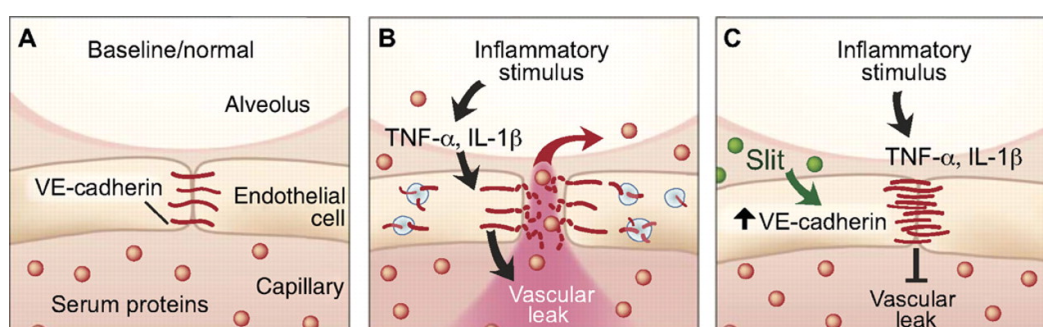


Fig. 4.19 Cartoon demonstrating the effect of SLIT-2 ROBO-4 signalling on pulmonary endothelial leak. Slit-2 binds ROBO-4 receptors which inhibit VE-cadherin internalisation, maintaining the gap junctions between endothelial cells and preventing leak into the alveolar space.

Figure taken, with permission, from London et al (2010)¹⁶⁹

Targeting Robo4-Dependent Slit Signalling to Survive the Cytokine Storm in Sepsis and Influenza

Nyall R. London, Weiquan Zhu, Fernando A. Bozza, Matthew C. P. Smith, Daniel M. Greif, Lise K. Sorensen, Luming Chen, Yuuki Kaminoh, Aubrey C. Chan, Samuel F. Passi, Craig W. Day, Dale L. Barnard, Guy A. Zimmerman, Mark A. Krasnow, Dean Y. Li

Publication: *Science Translational Medicine*

Publisher: *The American Association for the Advancement of Science*

Date: Mar 17, 2010

Copyright 2010, American Association for the Advancement of Science

Gene Name	Human Protein	Function (UniProt database)	Reference identifier (Pubmed ID)
<i>DEFA4</i>	Neutrophil defensin 4	Antimicrobial activity	15616305 ¹⁷²
<i>LCN2</i> **	Neutrophil gelatinase-associated lipocalin (NGAL)	Limits bacterial proliferation by sequestering iron bound to microbial siderophores, such as enterobactin	28214071 ¹⁷³
<i>CTSG</i>	Cathepsin G	Cleaves complement C3. Has antibacterial activity against <i>Pseudomonas aeruginosa</i>	1937776 ¹⁷⁴
<i>BPI</i> **	Bactericidal permeability-increasing protein	Cytotoxic to Gram-negative bacteria	2722846 ¹⁷⁵
<i>PRTN3</i>	Proteinase-3 (PR3)	Serine protease that degrades extracellular protein. Major component of neutrophil azurophilic granules	22266279 ¹⁷⁶
<i>RNASE3</i>	Eosinophil cationic protein	Exhibits antibacterial activity, including cytoplasmic membrane depolarization of preferentially Gram-negative, but also Gram-positive strains.	2501794 ¹⁷⁷
<i>AZU1</i>	Azurocidin	Neutrophil granule-derived antibacterial glycoprotein	2312733 ¹⁷⁸
<i>PGLYRP1</i>	Peptidoglycan recognition protein 1	Pattern receptor that binds to murein peptidoglycans (PGN) of Gram-positive bacteria.	11461926 ¹⁷⁹
<i>ELANE</i>	Neutrophil elastase	Digests outer membrane protein A in <i>E.coli</i> and <i>K.pneumoniae</i>	10947984 ¹⁸⁰
<i>LTF</i>	Lactotransferrin	Binds to the bacterial surface and is crucial for the bactericidal functions of other proteins	1599934 ¹⁸¹

Table 4.2 Transcripts from the midnight blue gene module that enriched for the pathway ‘antimicrobial humoral response. ***BPI* and *LCN2* have previously been identified by Kangaris *et al* (2015) as having increased expression in patients ARDS, which they confirmed by using qPCR.⁶³

4.3.10 Regulation of expression of SLITs and ROBOs discriminates the ‘blue’ and ‘grey’ MOSAIC clusters

An LDA model was fitted to differentiate the ‘blue’ and ‘grey’ MOSAIC clusters using samples from with respiratory SOFA score > 2. Mean accuracy for this fitted model was equal to 0.80 (95% CI 0.88-0.72) and AUROC was equal to 0.75. This had a lower accuracy compared with the fitted models for the other clusters. This suggested that the ‘grey’ and ‘blue’ clusters were more alike compared with the ‘red’ cluster as was apparent in Figure 4.16. This result was still consistent with the model being an accurate classifier given the small number of samples used to fit the model.

The discriminant coefficients were ranked by magnitude and these results for the ten highest ranking variable can be seen in Figure 4.20. The highest ranking immune mediator was IL-6. IL-6 concentrations in the ‘blue’ cluster were significantly higher than the ‘grey’ cluster (Figure 3.9 and Appendix Figure D.2).

The highest ranking gene module was again the ‘royal blue’ module which enriched for the pathway ‘regulation of SLITs and ROBOs’ (R-HSA-9010553, adjusted $p = 6.3 \times 10^{-4}$). This strongly implicated the role of this pathway in the mechanisms responsible for the ‘blue’ cluster given it was also a strong discriminator between the ‘red’ and ‘blue’ MOSAIC clusters.

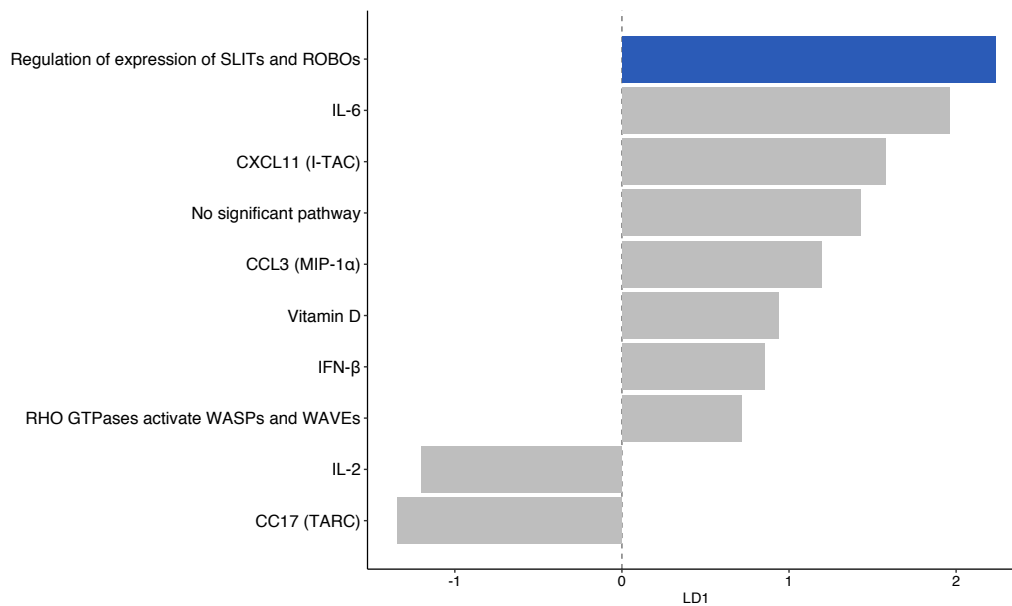


Fig. 4.20 Top ten ranked discriminators for samples with respiratory SOFA > 2 from the MOSAIC study between ‘blue’ and ‘grey’ clusters. These discriminators were selected by their relative magnitude. The top ranking gene module that was associated with a biological process is highlighted in blue (‘regulation of expression of SLITs and ROBOs’).

A line plot showing all of the ranked variables, ordered by effect size can be seen in Appendix Figure G.1.

4.3.11 Linear discriminant analysis of HARP-2 clusters

The HARP-2 investigators did not collect samples for measurement of gene expression. Clusters were, therefore, defined by the protein biomarker concentrations alone. A multi-class LDA model was fitted to determine how separated these clusters were. Mean accuracy for this model was equal to 0.81 (95% CI 0.73-0.88), multi-class AUROC was equal to 0.83. The model performed well at predicting minority class instances (mean accuracy = 0.86). The coefficients from this fitted LDA model were used to transform data points to project them onto linear discriminant axes, as shown in Figure 4.21.

Using only six biomarkers, it was apparent that there was less linear separation between the clusters when compared with LDA projection of protein biomarker values from patients recruited to the MOSAIC and GAIN studies.

The values of the linear discriminant coefficients are presented in Table 4.3. IL-6 and MMP-8 strongly discriminated the ‘dark red’ cluster (1) from the ‘dark green’ cluster (3). sRAGE and Ang-2 strongly discriminated the ‘dark yellow’ cluster (2) from the ‘dark green’ cluster (3). Finally, sRAGE, SP-D and MMP-8 strongly discriminated the ‘dark yellow’ (2) and ‘dark red’ (1) clusters.

	Dark green - Dark red	Dark yellow - Dark green	Dark yellow - Dark red
SP-D	-0.2198	0.1697	-0.5176
sTNFR-1	-0.3599	-0.1236	-0.0901
sRAGE	-0.4682	-1.3647	1.1324
MMP-8	-0.5356	0.0881	-0.4987
IL-6	-0.7771	-0.3961	0.0280
Ang-2	-0.4609	-0.4256	0.2492

Table 4.3 Ranked linear discriminator coefficients for each pairwise endotype comparison. The greater the magnitude (absolute value) of the discriminator the more important the variable. For example, sRAGE strongly discriminated the ‘dark yellow’ endotype from the other two endotypes.

By elimination and comparison of the protein biomarker concentrations in each cluster (Figure 3.12 and Appendix Figure D.3) it was confirmed that the discriminant characteristics of each of the three HARP-2 clusters could be termed dark red: MMP-8 driven , dark yellow: sRAGE driven , dark green: hypo-inflammatory .

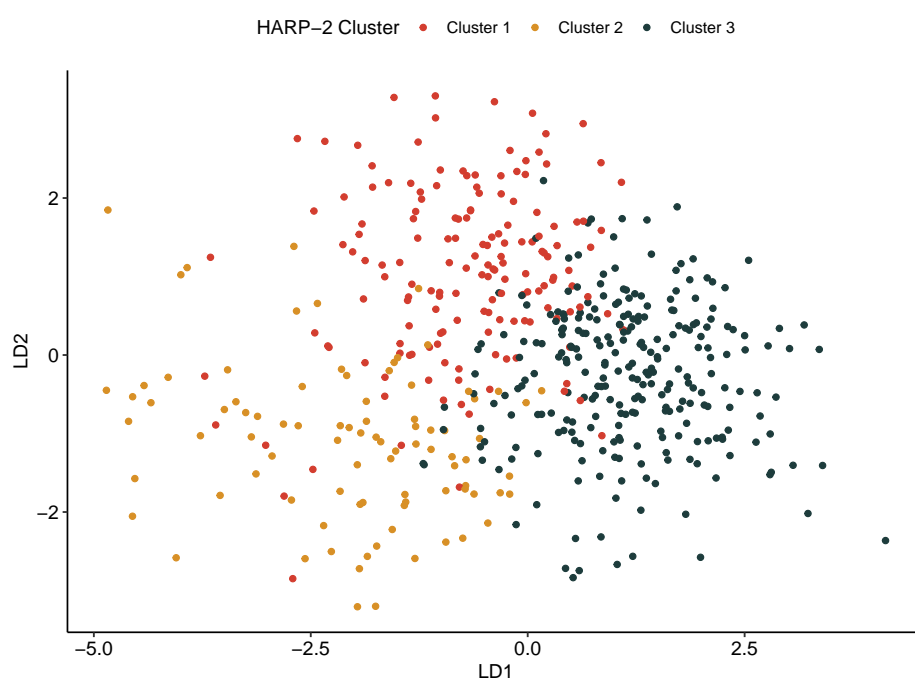


Fig. 4.21 LDA projection of the protein biomarker data from the HARP-2 study. The three clusters were defined using hierarchical clustering. The clusters shown here were less separated than has been demonstrated in the other two studies (MOSAIC and GAINs). The clusters here were determined using six protein biomarkers, four of which were not measured in patients recruited to the GAINs or MOSAIC studies.

4.4 Summary and discussion: integration of biological data

Integration of gene expression and other results from high throughput biological experiments is challenging because the number of measured gene probes in a standard gene expression experiment exceeds all other clinical or biomarker data many fold. In this section two different approaches have been taken to integrate these data types: differential gene expression between clusters and linear discriminant analysis using protein biomarkers and the explained variance of gene modules.

4.4.1 Differential gene expression

Differential gene expression demonstrated directional changes in gene expression enriched for plausible mechanisms between clusters found in the MOSAIC study (Figure 4.2). The ‘grey’ cluster was characterised by lymphocyte activation, adaptive immunity and haemostasis. The transcripts that enriched for the haemostasis pathway related to recovery and restorative processes. This cluster had lymphocyte counts that were no different to the other clusters, so it was termed ‘adaptive’. IFN- α 2a and CCL5 (RANTES) concentrations were significantly higher in this cluster than the other two. This suggested an interferon and activated lymphocyte-driven immune response which was consistent the enrichment of gene expression results here.

The MOSAIC ‘red’ cluster was differentiated from the MOSAIC ‘grey’ cluster by transcripts associated with neutrophil activation. The ‘red’ cluster had significantly higher concentrations of TNF- α , IL-6 and IL-8 which was consistent with this expression profile.

The ‘blue’ cluster was differentiated from the ‘grey’ cluster by transcripts that were associated with neutrophils (*RETN* and *CD177*), but they did not enrich for a known biological process or pathway. Surprisingly there were no differentially expressed genes between the ‘red’ and ‘blue’ clusters (Figure 4.2). These two groups had quite different immune mediator profiles so this was unexpected. Possible explanations for this could be related to the ‘blue’ cluster being a less severe phenotype than the ‘red’ but with the same underlying processes. Alternatively, the clustering method uncovered another layer of heterogeneity that was not captured by the analysis of immune mediator concentrations in isolation. Differentiation of these two clusters would therefore require additional variables of classification methods.

Of additional note was the finding that three transcripts differentiated both the ‘blue’ and ‘red’ clusters from the ‘grey’ cluster: *RETN*, *ZDHCC19*, *CD177*. *RETN* codes of resistin which is associated with insulin resistance but also plays a role in neutrophil degranulation.

CD177 is expressed on neutrophil cell membranes, binds PECAM-1 and may have a role in the transmigration of neutrophils. CD177 has recently been implicated as a predictor of poor outcomes in influenza patients.¹⁸² The authors of this study also used a WGNCA-based approach to gene expression analysis in patients with influenza.

ZDHHC19 encodes a palmitoyltransferase enzyme which palmitoylates protein residues with the aspartate-histidine-histidine-cysteine motif (DHHC). Its function is poorly characterised in the literature but it may play a role in modulating the function of STAT3, an important signalling protein in inflammation and immunity. *ZDHHC19* is one of six genes that were identified by the GAINs group in predicting the SRS phenotype in faeculent peritonitis.⁹⁸

There was no significant differential gene expression for pairwise comparisons between the three clusters found in the GAINs study. This was unexpected as there were pronounced differences between the protein biomarker concentrations in each cluster. The profiles of protein biomarkers in the GAINs clusters were more polarised compared with the clusters identified in the MOSAIC study.

There are two possible explanations for these results. The most plausible explanation is that there was additional heterogeneity between patients in each cluster which was not accounted for by protein biomarker concentrations or gene expression. Additional information would therefore be required to discriminate these data points and attribute biological mechanisms.

Another possible explanation could be related to the pre-processing and batch correction methods which may have suppressed the differences between clusters. The pre-processing was conducted using the same commands as for the MOSAIC data. Batch correction modifies probe intensity values and this may have caused excessive adjustment of gene probe variances. The batch effect correction function used in this analysis (ComBat) is the same function that was used by the GAINs research group in their publications and is widely used in the literature.^{97,183}

4.4.2 Correlation between WGCNA-derived gene modules and clusters

There were no statistically significant correlations between gene modules and clusters from the GAINs study. Nor did any clinical characteristics correlate with gene modules in these patients. Two possible explanations include unaccounted heterogeneity between patients in each cluster or that the *consensusBlockwiseModules* WGCNA function had over-suppressed

the variances of transcript levels when constructing the coexpression network across four different microarray experiments.

The second reason might have been addressed by the derivation of gene networks using only one microarray experiment at a time. This approach would generate four sets of gene modules. Correlations with cluster assignments and clinical characteristics could be calculated. However each set of gene modules could not be compared directly and inferences would be based on smaller samples sizes and increase the number of statistical comparisons. Results may not be consistent across each of the four networks and so these findings would be difficult to interpret.

The gene modules identified in the MOSAIC gene expression data showed strong correlations between many clinically relevant features. This analysis identified a gene module that enriched for the process ‘antimicrobial humoral response’ that was significantly correlated with poor patient outcomes. Secondary bacterial infection of patients with influenza is a recognised complication that is associated with morbidity and mortality in these patients. Inferring bacterial infection using the highest-ranked enrichment process from an abstracted gene expression network is contentious without objective evidence of infection. The relationships between gene modules, clusters and clinical characteristics are explored in more detail in Chapter 5.

The correlation coefficients between the MOSAIC clusters and gene modules found that the ‘red’ module was positively correlated with a gene module that enriched for the GAIT mechanism. The ‘grey’ module was negatively correlated with this pathway, but this was not statistically significant after adjustment. This mechanism directly relates to IFN- γ modulation of the acute phase of the immune response and to anti-RNA virus cellular mechanisms. The GAIT mechanism is thought play an important role in monocytes.¹⁵⁰ Identification of this mechanism in this context led to further exploration of how IFN- γ plays an essential role in the immune response to viral infection and how influenza may be able to subvert this mechanism to avoid detection and cause harm to its host.

4.4.3 LDA of clusters

I used the downsampling and noise-reduction methods that WGCNA offered, to calculate a mathematical relationship between each sample and modules of highly connected genes. The explained variance values of each sample with gene modules enabled gene expression information to be incorporated with the protein biomarker concentrations without excessive expansion of the features for each sample.

Augmentation of data risks disruption of the existing relationships between data points by adding excessive noise to or changing the variance of the data. Examination of principal component projections of the combined eigengene and protein biomarker values showed that this augmentation was in an orthogonal direction to the protein biomarker values (Figure 4.9 and Figure 4.10). This meant that the explained variance attributed to protein biomarker concentrations in each sample was preserved.

Use of linear discriminators showed how effective they were at projecting data with linear transformations into easily separable groups (Figures 4.11 and 4.16). Decision boundaries between clusters might have been calculated using other classification algorithms (random forests, decision trees, support vector machines). These other classification methods are designed for binary (two-group) classification. Furthermore, once these alternative methods of classification are fitted, establishing the each contributing variable's importance is not always transparent and may involve removing variables and repeatedly refitting models. LDA possibly offered poorer classification performance than other methods. However, application of the LDA method was straightforward when making inferences about the data. In the models that were fitted, LDA accuracy and AUROC statistics were generally robust despite the relatively small sample sizes in each cluster.

Using the LDA model's scaling factors as feature importance measures helped to determine the key mechanisms in each cluster. The identified mechanisms were concordant with protein biomarker concentrations and credible.

These methods identified two plausible pathways in the samples from the GAINs study. The neutrophil activation pathway discriminated the 'purple' and 'green' clusters, which was consistent with their protein biomarker profiles. Metabase sub-network analysis of the combined 'black' and 'dark orange' modules showed that key genes involved in secretion of cytokines, which are known to be implicated in familial HLH. The pattern of cytokine dysregulation in the 'purple' cluster appeared to be extreme as concentrations of both pro- and anti-inflammatory cytokines were raised. An HLH-like syndrome might have been the cause of this phenotype. Enrichment of the sub-network that included RIPK3 was also of interest. Involvement of the inflammasome and necroptosis processes in these samples could be verified by measurement of serum IL-18 and DAMPs associated with necroptosis.¹⁸⁴ Similarly, HLH might have been confirmed by measurement of ferritin, triglycerides and soluble IL-2 receptor (CD25).¹⁶³

A valid criticism of the approach taken here was the decision to combine the 'black' and 'dark orange' modules to arrive at these results. These two gene modules were closely related

according to the gene module dendrogram (Figure 3.18), and both highly ranked in the LDA model and so this step was considered reasonable.

‘Regulation of protein phosphorylation’ (GO:0001934) discriminated the ‘green’ and ‘yellow’ GAIN clusters. This pathway contained five transcripts that were all associated with immune and vascular themes. These transcripts did not directly relate to neutrophils and suggested that a different set of inflammatory processes characterised the immune response in these clusters. The lack of directionality of LDA coefficients meant it could not be inferred which cluster was associated with these transcripts. Corroborating information from other features would also be required to determine the role of these genes.

Based on these results, the clusters from the GAIN study were termed as shown in Figure 4.22. These cluster labels are based on the protein biomarker profiles given the limited insight into the explanatory mechanisms that might underlie the ‘green’ and ‘yellow’ clusters.

The approaches taken in this chapter demonstrated plausible mechanisms that delineated each of the clusters in the samples from the MOSAIC study. This plausibility is based on corroboration of identified mechanisms. For example, the differential gene expression between the ‘red’ and ‘grey’ clusters identified transcripts that were up-regulated with respect to the ‘red’ cluster. These up-regulated transcripts enriched for neutrophil degranulation processes. LDA models confirmed that a neutrophil degranulation process discriminated these two clusters.

Similarly, the ‘regulation of SLITs and ROBOs’ pathway discriminated ‘blue’-‘grey’ and ‘blue’-‘red’ clusters. This strongly implicated the role of this pathway in patients in the ‘blue’ MOSAIC cluster.

Comparison of the ‘grey’ and ‘red’ MOSAIC clusters identified up-regulated transcripts in the ‘grey’ cluster that enriched for processes associated with lymphocyte activation and adaptive immunity. These pathways were consistent with the immune mediator profiles of the ‘grey’ cluster samples. There were no significant differences in lymphocyte counts between clusters (Figure 4.3) and so this cluster was termed ‘adaptive’.

A summary of the progress in delineating endotypes of ARDS and severe influenza based on protein biomarker and transcriptomic analysis is shown in Figure 4.22.

Given the identification of plausible mechanisms for some of the clusters the term ‘endotype’ will now be used for greater clarity, as the focus of this thesis pivots from endotype discovery to endotype characterisation.

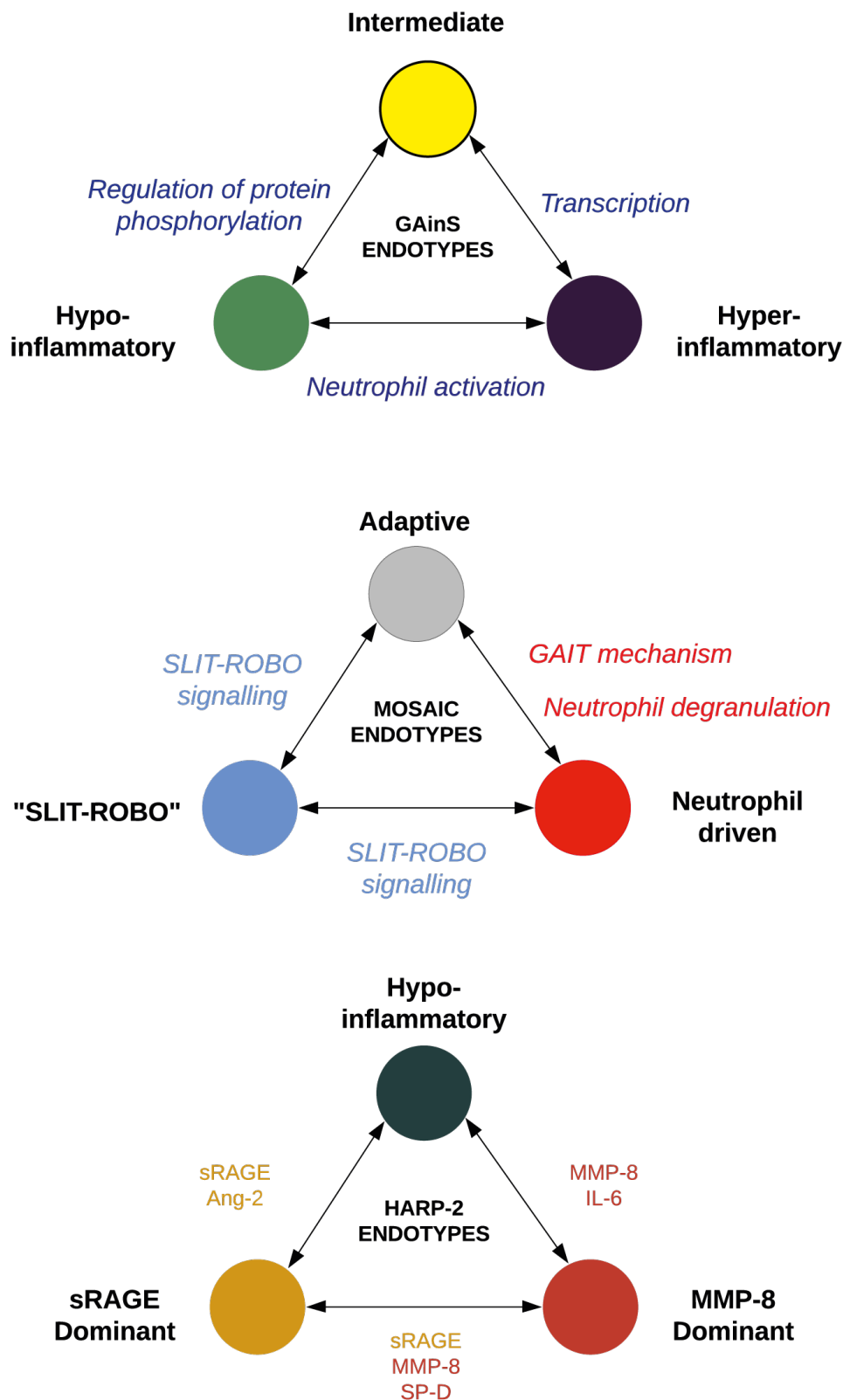


Fig. 4.22 Representation of endotypes based on cytokine clustering, differentially expressed genes and ranked linear discriminators. This is a preliminary representation of the endotypes as it does take into account clinical information, but serves as a summary of the progress in delineating subtypes of ARDS and severe influenza.

CHAPTER 5

Endotype characterisation

Chapter 4 demonstrated how transcriptomic data could be used to describe the mechanisms that might underlie different immune profiles of ARDS or severe influenza infection. This chapter will seek to determine if clinical variables can further characterise these endotypes and if they are associated with different patient outcomes. Where there were multiple sampling times for protein biomarkers, endotype stability will also be explored.

The clusters identified in the HARP-2 study will be characterised based on protein biomarkers and clinical features alone.

Statistical comparisons between groups were made using either ANOVA with Tukey's *post hoc* test for normally distributed variables or Kruskal-Wallis test with Dunn's test for non-normal variables. Ordinal variables, for example, the cardiovascular component of the SOFA score, were compared using the Kruskal-Wallis test. Logistic regression was used for binary variables.

5.1 GAinS endotypes

5.1.1 With the exception of PaO₂-FiO₂ ratio, there were no significant differences in organ dysfunction between the endotypes in the GAinS study

We compared the clinical features of patients in each endotype to determine clinical phenotypes. There were 196 samples with protein biomarker and transcriptomic data from 144 individual patients. 71 samples were from 55 patients with ARDS. To characterise each endotype comparisons were made irrespective of ARDS status in order to determine the

broader clinical features associated with each endotype. Clinical variables from the subset of patients with ARDS were also compared.

Neutrophil counts were not recorded in the GAINs study. The total white cell count was not significantly different between the GAINs endotypes, irrespective of ARDS status (Table 5.1, Figure 5.1 and Figure 5.2). Given the protein biomarker profile and gene module associations of the hyper-inflammatory endotype, significant differences in total white cell count might have been expected, especially between the hypo-inflammatory and hyper-inflammatory endotypes. This was not shown to be the case. Total white cell count did not differentiate clusters in the GAINs study.

The hyper-inflammatory endotype would have been expected to have had worse multi-organ dysfunction compared with the other endotypes. Compared with the hypo-inflammatory cluster the levels of creatinine, bicarbonate, requirement for high dose vasopressor support, renal replacement therapy and platelet count all appear to be consistent with worse multi-organ dysfunction (Table 5.1). These differences were statistically significant when considered in isolation but not after adjustment. A larger sample size might have demonstrated these differences better.

The intermediate endotype had a very similar clinical phenotype to the hyper-inflammatory cluster, despite their very different protein biomarker profiles. One statistically significant difference was the lower PaO₂-FiO₂ ratio for patients in the hyper-inflammatory endotype compared with the patients in the intermediate endotype (Dunn's test, $Z = 3.05$, adjusted $p = 0.007$).

5.1.2 Patients in the GAINs study with the hypo-inflammatory endotype were more likely to have received steroid therapy

Patients with the hypo-inflammatory endotype were significantly more likely to have received steroid therapy compared with the intermediate endotype (OR = 3.5, 95% CI 1.3-9.8, $p = 0.02$, Table 5.1). This might have been a plausible explanation for the globally suppressed cytokine profile in the hypo-inflammatory patients, although no difference in steroid use was apparent in the hyper-inflammatory endotype. This finding recapitulates the enriched sub-networks identified in the combined 'black' and 'dark orange' gene modules using the Metabase tool. Sub-network (v.) in Figure 4.15 contained two transcripts related to glucocorticoid receptors (*FKBP5* and *SGK1*). These two gene modules were discriminant between the hypo-inflammatory and hyper-inflammatory endotypes, not the intermediate endotype. This does not quite fit with the relationship identified in the clinical data, although

the globally depressed cytokine concentrations in these patients is consistent with steroid therapy.

	All patients			Patients with ARDS		
	Hypo-inflammatory	Hyper-inflammatory	Intermediate	Hypo-inflammatory	Hyper-inflammatory	Intermediate
n	27	47	70	8	24	23
Sex = Female n (%)	12 (44%)	14 (30%)	29 (41%)	6 (75%)	10 (42%)	8 (35%)
Age mean (sd)	67.4 (15.1)	68.2 (11)	64.2 (15.8)	63.4 (16)	68.5 (12.1)	65.5 (13.9)
Community acquired pneumonia	17 (64%)	17 (72%)	40 (57%)	5 (40%)	19 (79%)	17 (74%)
Faeculent peritonitis	10 (37%)	13 (28%)	30 (43%)	3 (60%)	5 (21%)	6 (26%)
Charlson comorbidity index (median)	1	1	1	1	1	1
Steroid treatment	10 (58.8%)	13 (27.7%)	10 (14.3%)	2 (25%)	5 (20.1%)	3 (13%)
White cell count (x10 ⁹ /mL) mean (sd)	16.8 (8.7)	18.7 (9.7)	18.6 (8.2)	17.6 (6)	17.9 (7.7)	16.7 (7.2)
Platelet count (x10 ⁹ /mL) mean (sd)	174 (72)	181 (91)	184 (108)	164 (55)	194 (100)	205 (136)
Creatinine (umol/L) median (IQR)	102 (69-124)	107 (81-230)	115 (78-174)	104 (74-126)	108 (80-215)	116 (75-140)
Bilirubin (umol/L) median (IQR)	14 (8-23)	14 (11-28)	17 (10-32)	21 (11-28)	13 (11-23)	16 (9-34)
Bicarbonate (mmol/L) mean (sd)	21.8 (9.3)	20.6 (4.8)	21.5 (5.8)	21 (9.5)	21.8 (4.9)	22.7 (5.1)
Lowest PaO ₂ -FiO ₂ ratio median (IQR)	16.5 (12.4-22.8)	11 (8.7-20.2)	20 (11.4-27.7)	12.3 (9.4-19.8)	12.1 (9.5-20.8)	18.4 (12.7-22.9)
Cardiovascular SOFA >2 n (%)	12 (44%)	30 (63.8%)	39 (55.7%)	3 (37.5%)	15 (62.5%)	11 (47.8%)
Renal replacement therapy n (%)	2 (7.4%)	11 (23%)	10 (14.3%)	1 (12.5%)	5 (20.8%)	2 (8.7%)
Length of stay (survivors) median days (IQR)	23.5 (18.5-32.5)	20 (10-40)	21 (12-33)	29.5 (21.3-55)	27 (13-35)	21 (8-34)
In hospital mortality n (%)	7 (25.9%)	18 (38.3%)	21 (30%)	2 (25%)	11 (45.8%)	10 (43.5%)
30 day mortality , unadjusted HR (95% CI)	1 (reference)	2.05 (0.66-6.36)	1.84 (0.61-5.5)	1 (reference)	2.69 (0.33-22)	4.2 (0.52-33)

Table 5.1 Patient characteristics in each of the three endotypes identified in the GAINs study. There were two statistically significant differences between endotypes (denoted in bold): 1. PaO₂-FiO₂ ratio was lower in patients with the hyper-inflammatory endotype compared with the and intermediate endotype. 2. Patients with the hypo-inflammatory endotype were more likely to have received steroids compared with the intermediate endotype.

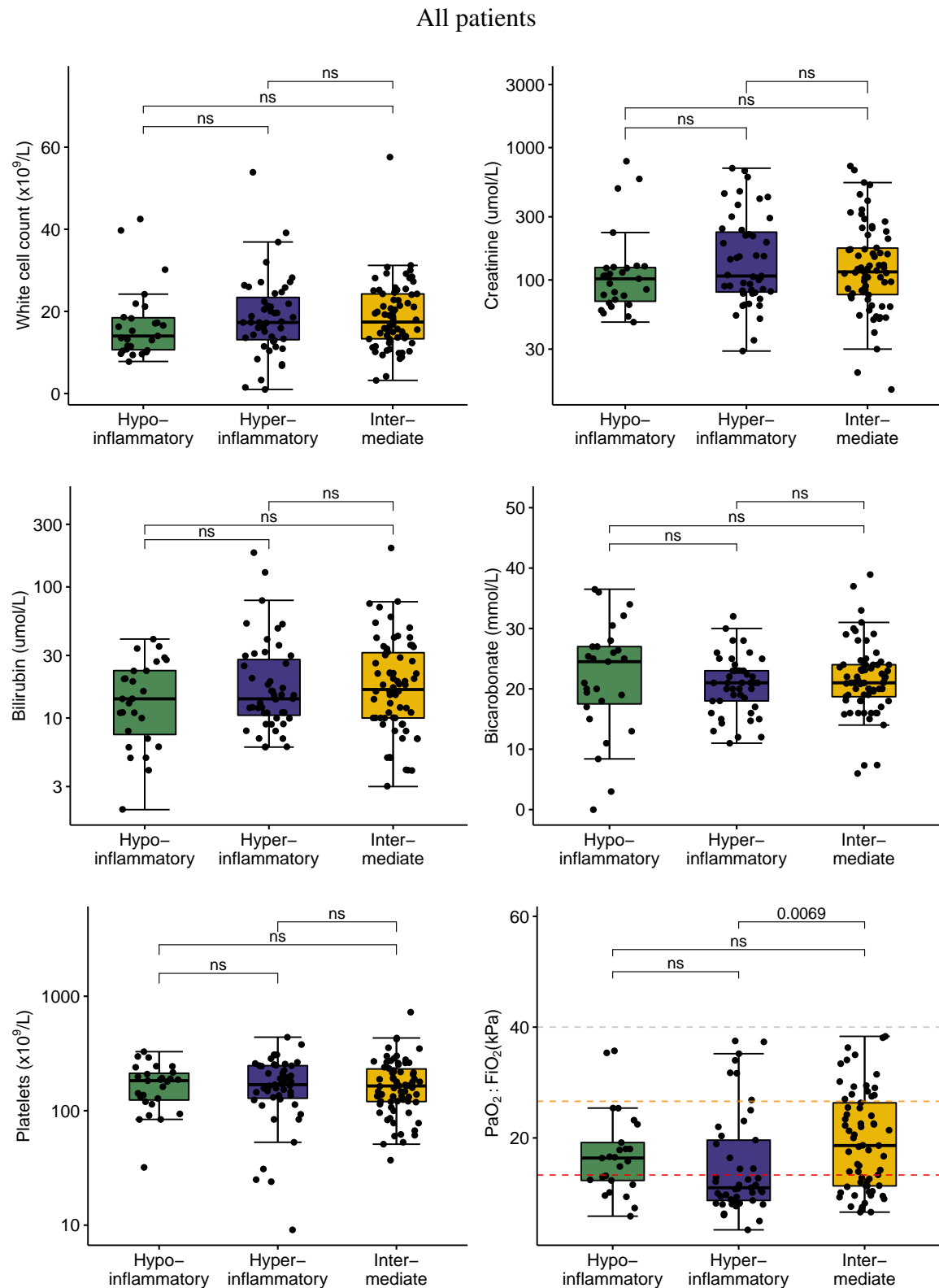


Fig. 5.1 Boxplots comparing the differences in clinical variable measurements related to organ dysfunction between patients, *with or without ARDS*, from each endotype identified in the GAINs study. p values are corrected for multiple comparisons. The dashed lines on the $PaO_2 : FiO_2$ plot show the thresholds for mild (<40), moderate (<26.6) and severe (<13.3) ARDS.

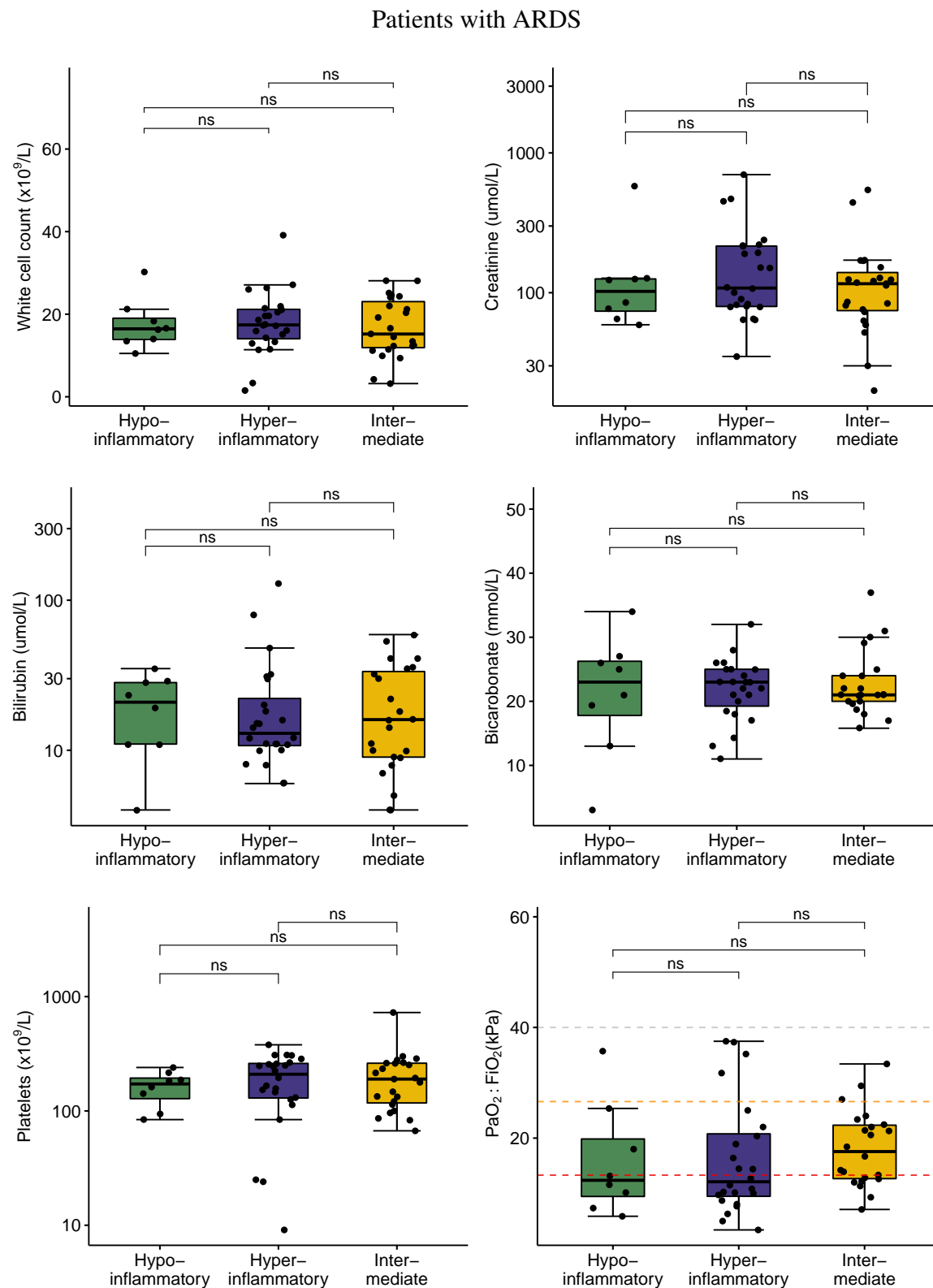


Fig. 5.2 Boxplots comparing the differences in clinical variables measurements related to organ dysfunction between patients *with* ARDS from each endotype identified in the GAINs study. *p* values are corrected for multiple comparisons. The dashed lines on the $PaO_2 : FiO_2$ plot show the thresholds for mild (<40), moderate (<26.6) and severe (<13.3) ARDS.

5.1.3 Survival analysis showed no significant differences between patients with different endotypes in the GAINs study

There were no differences between endotypes for in-hospital mortality, 30-day mortality or hospital length of stay, irrespective of ARDS status (Table 5.1). The Kaplan-Meier analysis of 30 day mortality for each endotype in all patients, and the sub-group with ARDS is shown in Figure 5.3.

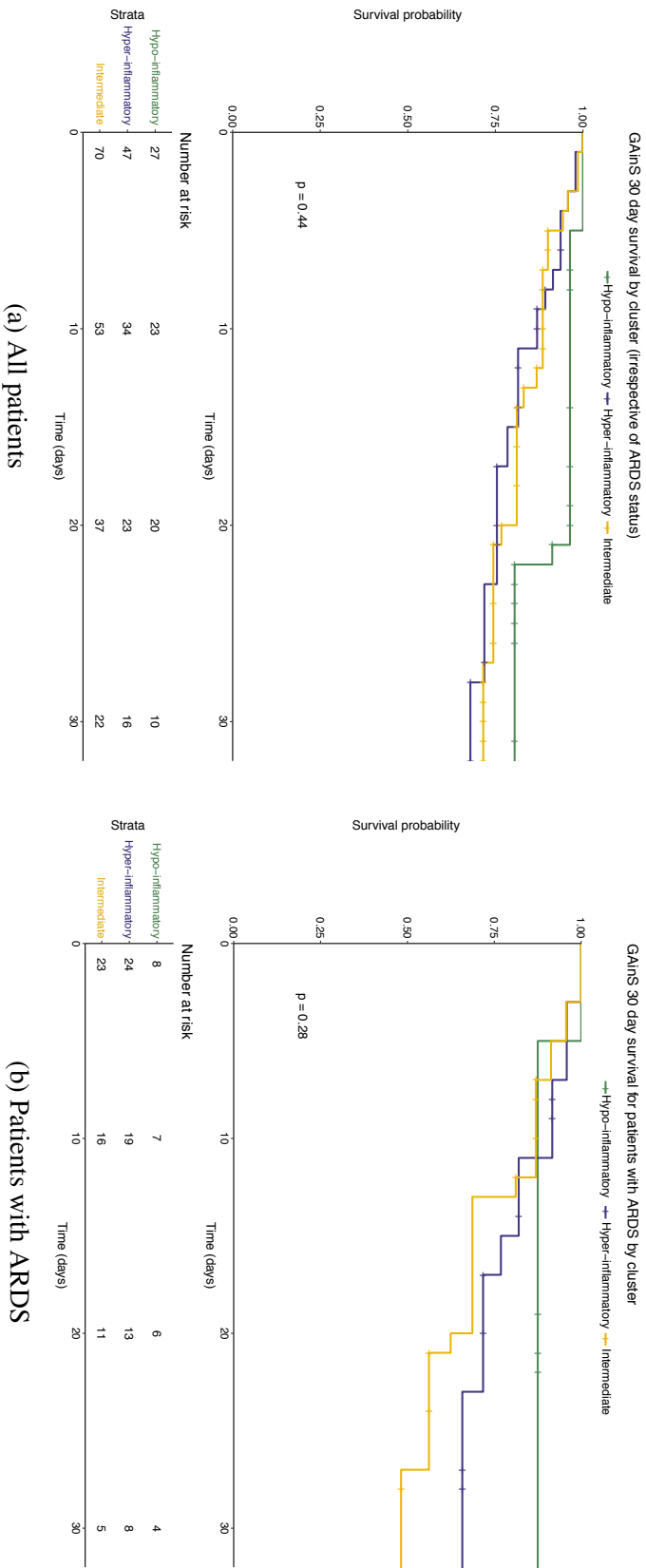


Fig. 5.3 Kaplan-Meier analysis of the 30 day mortality of all patients (a) and patients with ARDS (b), stratified by the endotypes identified in the GAiNS study. *p* values indicate the significance level of the log-rank test.

5.1.4 Endotypes identified in the GAINs study were not stable over measured time points

Hierarchical clustering of protein biomarkers measured in the GAINs study was conducted using the results from samples collected on study days 1, 3 and 5. This was to capture all possible immunological states using the protein biomarker data. The relative stability of endotypes and observation of their transitions was therefore observable. The Sankey diagram in Figure 5.4 shows these transitions alongside patient outcomes at these sampling times.

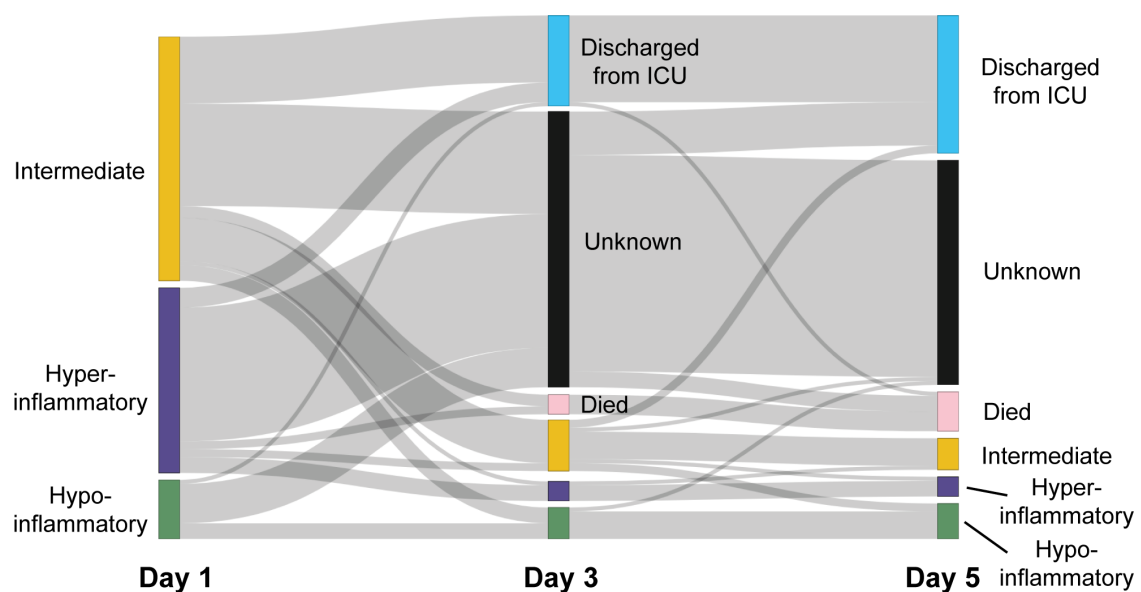


Fig. 5.4 Sankey diagram showing the transitions between endotypes on sampling days 1, 3 and day 5. Patient outcomes are also shown here. Although clusters were not particularly stable the transitions observed were plausible. For example, there were no transitions from one extreme endotype (hyper-inflammatory) to the other (hypo-inflammatory) and vice-versa.

The paucity of sampled biomarkers at later days 3 and 5 is apparent and limited the effectiveness of this analysis. 28 patients had measured biomarkers at day 1 and day 3. Of these, 24 also had results for biomarkers sampling at day 5.

The adjusted Rand index was calculated to determine cluster concordance at different sampling times as a measure of cluster stability. Between day 1 and day 3 it was equal to 0.28, and between day 3 and day 5 it was equal to 0.34. Both of these values suggested that cluster stability was low although this might be expected with small sample sizes.

The endotype transitions shown in Figure 5.4 appeared to be plausible. No transitions from a hyper-inflammatory to a hypo-inflammatory endotype (or the converse) were observed. Five patients transitioned from the intermediate endotype to the hypo-inflammatory endotype at day three. Of these three, two patients had received steroids. Three patients from the intermediate endotype at day three transitioned to the hypo-inflammatory endotype at day five. One of these patients received steroids. The relative incidence of steroid use in these patients was insufficient to explain the mechanisms relating to this endotype based on this information.

5.2 MOSAIC endotypes

5.2.1 Patients with the neutrophil driven endotype from the MOSAIC study had significantly worse multi-organ dysfunction.

Clinical features of endotypes were compared to determine whether the underlying mechanisms translated into plausible clinical phenotypes. For the MOSAIC study, the depth to which this analysis might have been possible was limited by missing data annotations and discordance between clinical information and biological samples in the study database. Low data integrity and poor completion of critical illness fields were more apparent in the MOSAIC study than the GAINs and HARP-2 studies. This might have been due to the study contexts being different; critical and non-critical in-patients (MOSAIC) compared with critical care patients (GAINs, HARP-2). The variables that were least likely to be missing were routine laboratory investigations (full blood count, biochemistry).

Table 5.2 shows that baseline demographics were no different between patients in each endotype. Patients with the neutrophil driven endotype were associated with lower platelet counts, higher creatinine and bilirubin levels compared with the adaptive endotype (Figure 5.5). Despite the polarised neutrophil and lymphocyte-like immune responses in the neutrophil drive and adaptive endotypes, there were no differences in the neutrophil and lymphocyte cell counts between these two groups of patients.

	Respiratory SOFA ≥ 2			Respiratory SOFA ≥ 3		
	Adaptive	SLIT-ROBO	Neutrophil driven	Adaptive	SLIT-ROBO	Neutrophil driven
N (at T1)	54	33	17	11	18	13
Demographics mean (sd)						
Male sex (%)	31 (57.4%)	18 (54.5%)	11 (64.7%)	6 (54.5%)	10 (55.6%)	9 (69.2%)
Age, years	45.2 (17.3)	48.5 (13.3)	44.7 (11.9)	42.7 (15.1)	44.8 (15.7)	42.8 (11.3)
BMI, kg m ⁻²	28.5 (8.5)	25.5 (4.3)	30.8 (8.7)	30.9 (8.9)	26 (4.6)	32 (9.3)
Past Medical History (n, %)						
Active smoker	20 (37%)	14 (42%)	9 (53%)	4 (36.4%)	9 (50%)	5 (38.5%)
Asthma	18 (33%)	7 (21%)	1 (5.9%)	1 (10%)	5 (27.8%)	1 (7.7%)
Other chronic lung disease	6 (11.1%)	6 (18.2%)	1 (5.9%)	2 (18.2%)	3 (16.7%)	1 (7.7%)
Cardiovascular disease	20 (37%)	12 (36.4%)	4 (23.5%)	4 (36.4%)	5 (27.8%)	3 (23.1%)
Diabetes	7 (13%)	2 (6.1%)	2 (11.8%)	2 (18.2%)	0	2 (15.3%)
Active malignancy	5 (9.3%)	5 (15.2%)	4 (23.5%)	1 (9.1%)	2 (11.8%)	2 (15.4%)
Clinical variables mean (sd) or median [IQR]						
Hb (g/L)	124 (25)	107 (22)	112 (18)	104 (27)	101 (17)	109 (14)
Neutrophil count ($\times 10^9/L$)	8.5 (4.9)	10.9 (7.8)	10.7 (9.4)	9 (4.7)	12.7 (8.6)	12 (9.9)
Lymphocytes ($\times 10^9/L$)	1.4 (0.99)	0.99 (0.47)	1.31 (2.7)	0.95 (0.46)	0.96 (0.39)	1.61 (3.04)
Platelet count ($\times 10^9/L$)	242 (114)	195 (94)	128 (95)	261 (119)	196 (103)	113 (70)
Creatinine (umol/L)	70 [56-85]	88.5 [67-176]	138 [78-278]	63 [43-91]	81 [54-166]	203 [129-294]
				p = 0.001	p = 0.003	p = 0.003
				p < 0.001	p = 0.01	p = 0.01
				(Both groups)	(Both groups)	(Both groups)
Bilirubin (umol/L)	9 [6-12]	12 [8-21]	19 [15-25]	9 [4-15]	12 [8-22]	16 [14-25]
Albumin (g/L)	30.3 (7.7)	22.2 (7.6)	20.5 (5)	24.4 (5.7)	18 (5.4)	19.5 (4.2)
				p < 0.0001	p = 0.01	p = 0.01

Table 5.2 Clinical characteristics of patients in each MOSAIC endotype. Only results from paired biochemical samples at recruitment (T1) are presented here. Comparison were made using ANOVA with Tukey's post hoc test. Values that were statistically significant are highlighted in bold with associated adjusted p values in the final column. The neutrophil driven endotype had evidence of multi-organ dysfunction. The SLIT-ROBO group was characterised by significantly lower albumin levels. These results can also be visualised as boxplots in Figures 5.5 and 5.6.

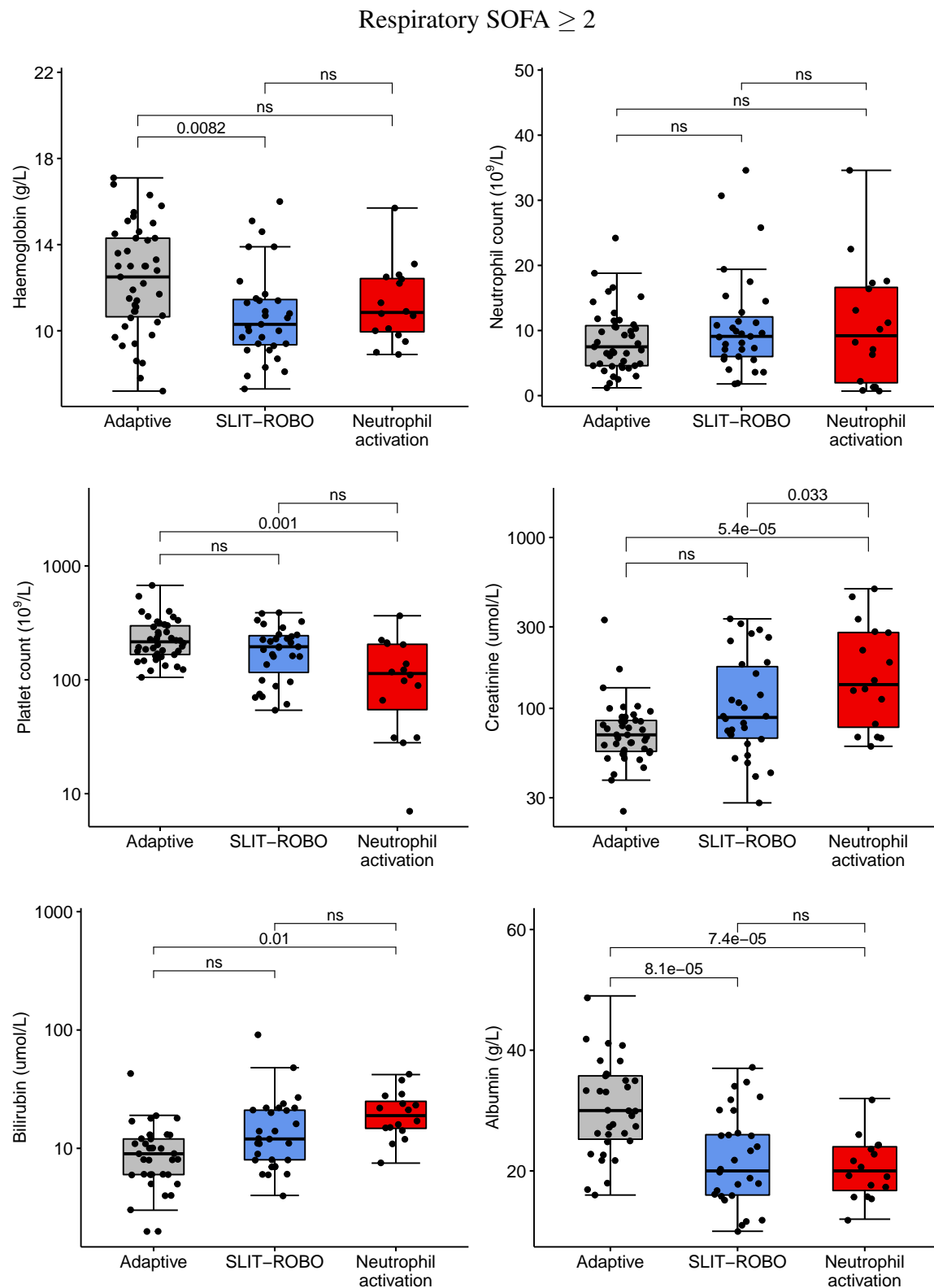


Fig. 5.5 Boxplots comparing clinical variable measurements related to organ dysfunction between patients, from each endotype identified in the MOSAIC study. p values were corrected for multiple comparisons.

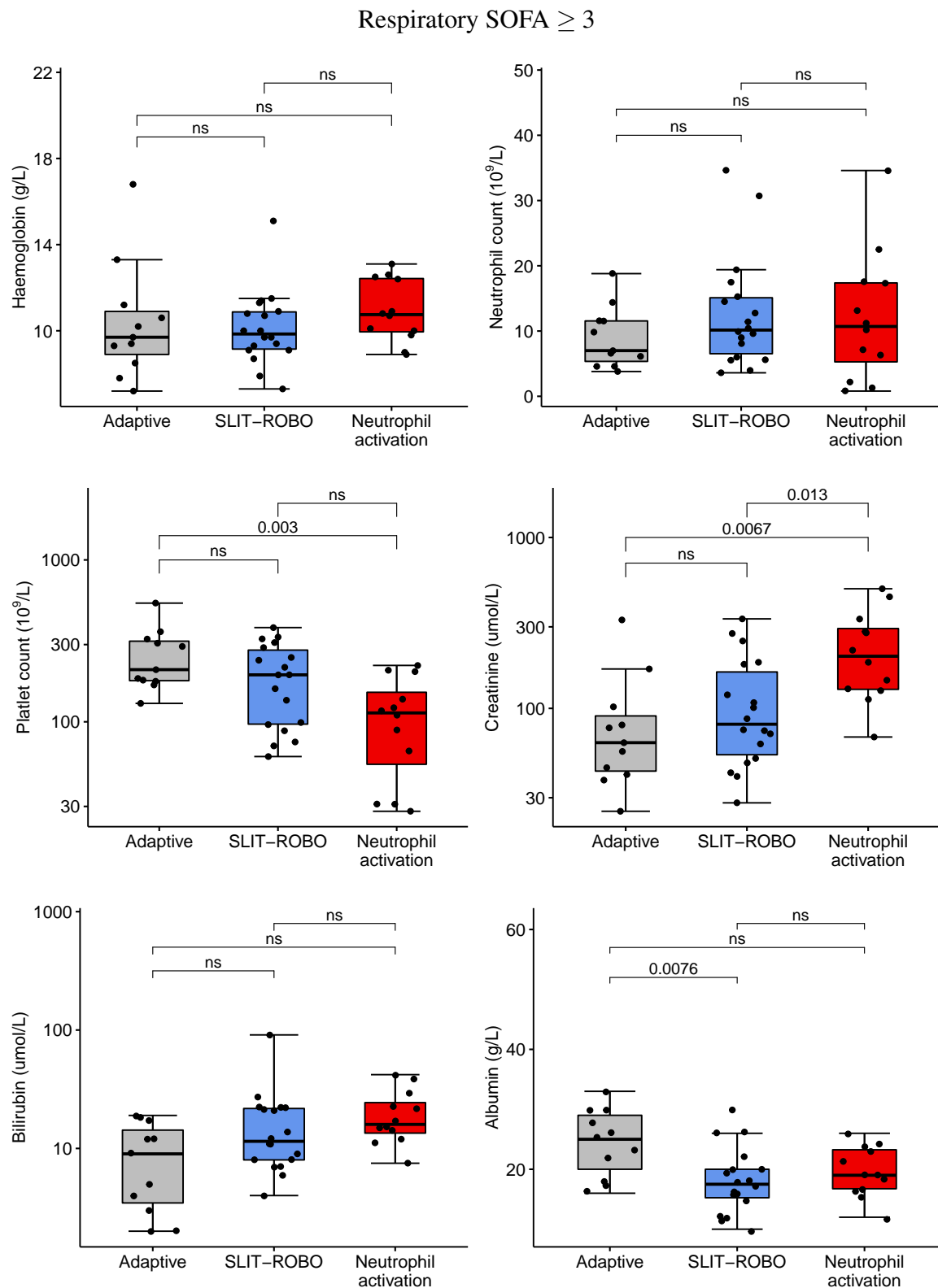


Fig. 5.6 Boxplots comparing clinical variable measurements related to organ dysfunction for patients with rSOFA ≥ 3 from each MOSAIC endotype. p values were corrected for multiple comparisons.

5.2.2 Low albumin concentration is associated with the SLIT-ROBO endotype

Patients with the SLIT-ROBO endotype were indistinguishable from the neutrophil driven endotype based on the clinical variables measured here. This was consistent with findings of the differential gene expression analysis between these two endotypes. These patients had significantly lower haemoglobin and albumin concentrations than patients with the adaptive endotype. These significant differences in albumin concentrations persisted in the smaller subset of patients with respiratory SOFA scores ≥ 3 (Figure 5.6).

Lower concentrations of both albumin and haemoglobin could be attributed to several causes: reduced synthesis, redistribution or increased consumption. Redistribution is probably the most likely cause in this context. Haemodilution from exogenous intravenous fluid administration or impaired regulation of capillary homeostasis are common in acute illness. In critical illness, loss of capillary homeostasis often manifests as apparent circulating hypovolaemia due to the loss of endothelial integrity, causing increased capillary leak.¹⁸⁵ Albumin and other plasma proteins leak into the tissues and measured circulating levels fall. The relative concentrations of these proteins are further reduced by administration of intravenous fluids, which are administered to restore intravascular volume. Given the underlying mechanism identified in patients with this endotype related to SLIT-ROBO signalling, which is implicated in endothelial integrity in animal models of sepsis and influenza infection,¹⁶⁹ this mechanism was considered a plausible cause of the low albumin and haemoglobin concentrations in these patients. Based on these results, the SLIT-ROBO endotype was termed **Endothelial Leak**.

5.2.3 MOSAIC endotypes are associated with different patient outcomes

Patients with the endothelial leak and neutrophil driven endotypes were more likely to be admitted to ICU, receive mechanical ventilation, require cardiovascular support and have a longer duration of hospital stay than patients with the adaptive endotype (Table 5.3). Mortality was significantly worse for patients with the neutrophil driven endotype whose unadjusted odds of hospital mortality compared with the adaptive endotype was equal to 11.1 (95% CI 2.8-44.6, $p < 0.001$). Identification of which patients received renal replacement therapy (RRT) was not possible as these fields were poorly annotated in the MOSAIC study database.

The critical care requirements for patients with the endothelial leak endotype were less but similar to the neutrophil driven endotype (48.5% compared with 64.7% requiring mechanical

ventilation). The 30-day mortality of patients in the endothelial leak endotype was however markedly different (Figure 5.7). Despite a significant proportion of patients with the endothelial leak endotype requiring treatment in critical care, their survival profile was similar to patients with the adaptive endotype, and significantly better than patients with the neutrophil driven endotype: 30-day mortality endothelial leak HR=0.79 (95% CI 0.118-5.72, $p = 0.8$) and neutrophil driven HR=5.1 (95% CI 1.04-25.6, $p=0.045$). Until this point in the analysis it had not been possible to distinguish between patients with the endothelial leak and neutrophil driven endotypes.

	Adaptive	Respiratory SOFA ≥ 2		Adaptive	Respiratory SOFA ≥ 3	
		Endothelial leak	Neutrophil driven		Endothelial leak	Neutrophil driven
N at T1	54	33	17	11	18	13
Adequately screened for evidence of bacterial infection, n (%)	25 (46%)	7 (21%)	12 (71%)	6 (54.5%)	7 (38.9%)	10 (76.9%)
Clinical evidence of bacterial infection, n (%)	11 (56%)	4 (57%)	10 (83%)	4 (36%)	3 (17%)	8 (62%)
	Reference	OR 0.59 (0.1-3.2) $p = 0.54$	OR 3.93 (0.71 - 21.7) $p = 0.12$	OR 2.7 (0.28-25.6) $p = 0.4$	Reference	OR 5.3 (0.62-46) $p = 0.12$
HDU/ICU Admission, n (%)	13 (24%)	20 (61%) OR 4.85 (1.94-13.7) $p < 0.001$	14 (82%) OR 14.7 (4.07 - 71.6) $p < 0.001$	-	-	-
Mechanical ventilation, n (%)	11 (20.4%)	16 (48.5%) OR 3.68 (1.44-9.8) $p = 0.007$	11 (64.7%) OR 7.2 (2.25-25.2) $p = 0.001$	-	-	-
Vasopressor support, n (%)	8 (14.8%)	17 (51.5%) OR 6.11 (2.2-16.9) $p < 0.001$	10 (58.8%) OR 8.2 (2.42-27.9) $p < 0.001$	8 (72.7%) NS	16 (88.9%) NS	10 (76.9%) NS
Hospital length of stay, median days [IQR]	6 [3.25 - 10.75]	15 (8-30) $p = 0.0002$	20 (8-29) $p = 0.004$	27 [8.5-31.5] NS	16.5 [11.5-36.5] NS	26 [19-34] NS
Hospital Mortality, n (%)	4 (7%)	7 (21%) OR 3.37 (0.9 - 12.6) $p = 0.07$	8 (47%) OR 11.1 (2.76 - 44.8) $p = 0.0007$	2 (18.2%) Reference	5 (27.8%) OR 1.73 (0.27-11) $p = 0.56$	6 (76.5%) OR 3.86 (0.59-25.3) $p = 0.16$
30 day mortality, n (%)	2 (3.7%)	2 (6.5%) OR 1.68 (0.22 - 12.5) $p = 0.61$	6 (35.3%) OR 14.2 (2.52 - 79.8) $p = 0.003$	2 (18.2%) Reference	2 (11.1%) OR 1.25 (0.1-15.6) $p = 0.86$	4 (15.4%) OR 4.44 (0.42-47.5) $p = 0.22$

Table 5.3 Outcomes for patients in each MOSAIC endotype. Categorical outcome variables were analysed with logistic regression using the adaptive endotype as the reference group. Mortality presented here is unadjusted for other variables. Length of stay was compared using a log-linear regression model. Values that were significantly different from the reference group were are highlighted in bold. SOFA: sequential organ failure assessment score. OR: odds ratio

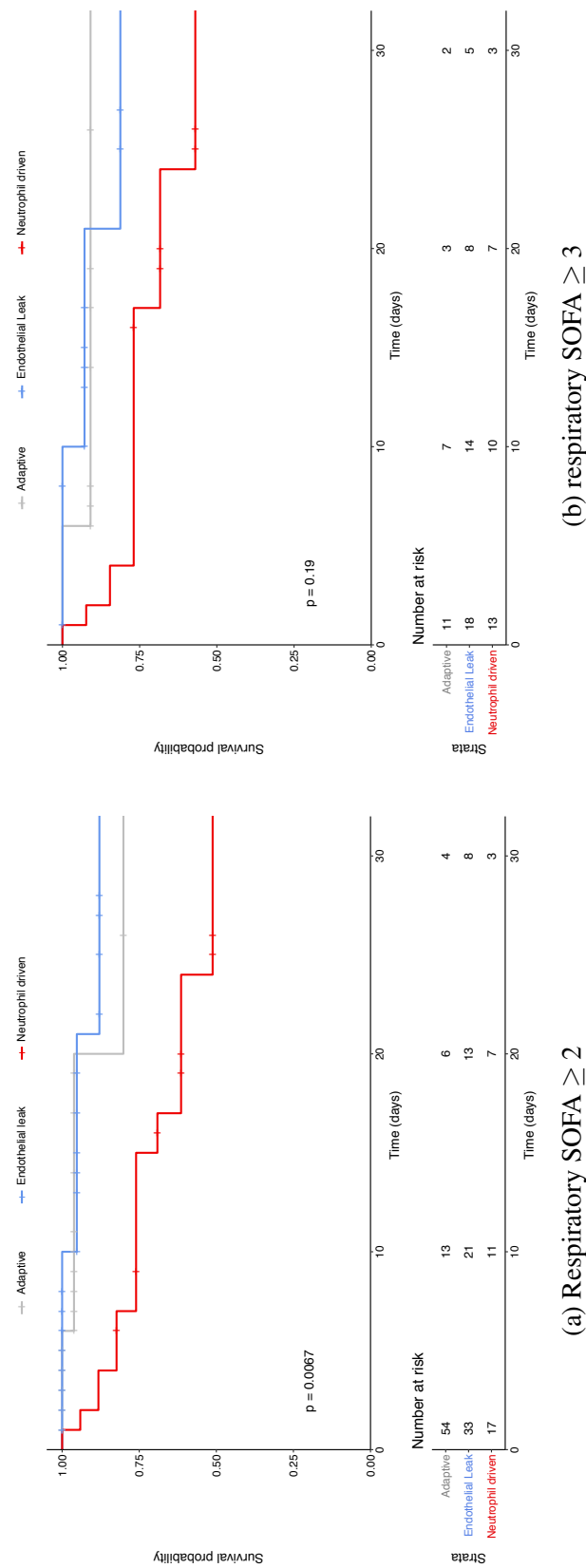


Fig. 5.7 Kaplan-Meier analysis of the 30 day mortality of patients with respiratory SOFA ≥ 2 (a) and respiratory SOFA ≥ 3 (b), stratified by MOSAIC endotype. p values indicate the significance level of the log-rank test.

5.2.4 MOSAIC endotypes were stable after 48 hours

The MOSAIC study investigators conducted biological sampling of patients at three time points: recruitment to the study (T1), two days afterwards (T2) and at least 7 days after discharge from hospital (T3). For this analysis, only T1 and T2 samples were analysed as the research question was to determine the endotypes of severe respiratory failure. T2 samples were taken at a median interval of 2.2 days after T1 samples.

There were 53 patients who underwent serum cytokine sampling at both T1 and T2. Of these only 5 patients (9.6%) transitioned from one endotype to another. The calculated adjusted Rank index was equal to 0.72 which is consistent with stable cluster assignments between T1 and T2. The cluster transitions and final hospital outcomes for patients can be seen in Figure 5.8.

Three patients (two from neutrophil driven , one from endothelial leak) transitioned to the adaptive endotype at T2. There were no transitions to the neutrophil driven endotype from the other two endotypes.

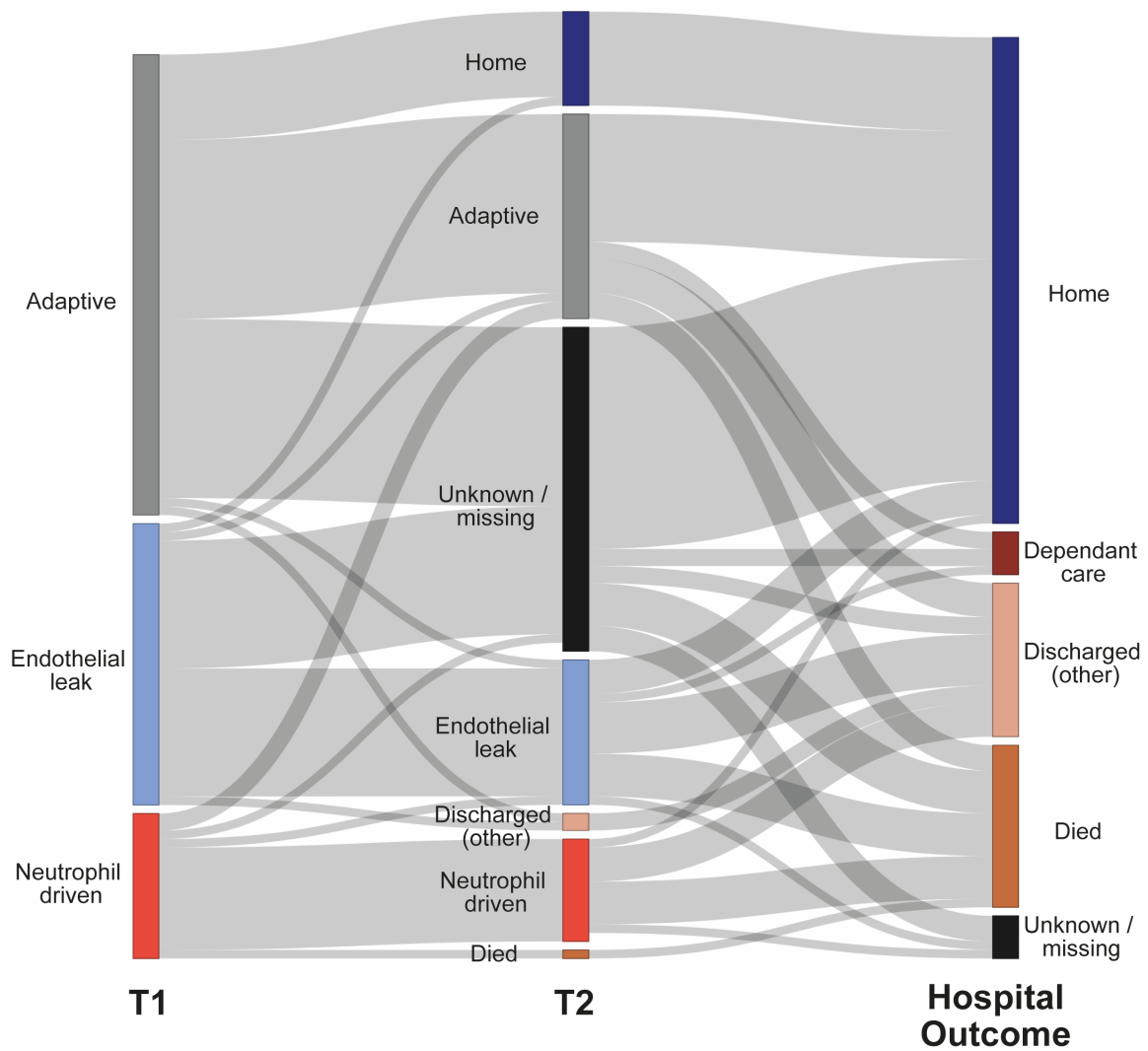


Fig. 5.8 Sankey diagram showing cluster transitions and hospital outcomes for patients within each endotype. The endotypes were generally stable at 48 hours. “Discharged (other)” refers to discharge to another secondary care institution (for example teaching to district general hospital).

5.2.5 The role of secondary bacterial infection in MOSAIC endotypes is uncertain

Secondary bacterial infection following influenza infection is widely recognised as a cause of morbidity and mortality.¹⁸⁶ Initial correlation analysis of gene modules with mortality indicated that the ‘antimicrobial humoral response’ module (Figure 3.20) was associated with mortality, suggesting that activation of immune pathways related to bacterial infection may be associated with worse outcomes.

The neutrophil driven endotype was associated with high levels of procalcitonin (Figure 3.12), and repeatedly associated with neutrophil activation and degranulation (Figures 4.17 and 4.2a). These associations suggested that secondary bacterial infection was a plausible explanation for why these patients were more unwell than patients in the other endotypes.

The MOSAIC investigators sought to address the question of secondary bacterial infection from the outset and carried out extensive sampling from recruited patients. Sampling methods to detect bacterial infection included:

1. Culture of all respiratory samples (nasopharyngeal aspirates, bronchoalveolar lavage, throat swabs, sputum).
2. Multiplex PCR quantification for *Staphylococcus aureus*, *Chlamydia pneumoniae*, *Haemophilus influenzae*, *Streptococcus pneumoniae*, *Pneumocystis pneumoniae*, *Legionella sp.*, *Klebsiella pneumoniae*, *Salmonella sp.*, *Moraxella catarrhalis* and *Bordetella pertussis* species in the above samples.
3. Quantification of bacterial 16S ribosomal RNA (rRNA) in all of the above respiratory samples, which measured the bacterial load.
4. Pneumococcal urinary antigen detection.
5. Results from routine microbiological investigations performed part of their routine care by treating medical teams. For example, blood, sputum, cerebrospinal fluid culture and urinary antigen test results.

Each case of suspected bacterial infection was discussed by a clinical panel following a review of all of the above results. The authors of the primary publication of the MOSAIC concluded found that severe disease was associated with gene expression modules that related to neutrophil activation, but they found no evidence of increased bacterial infection in these patients.¹¹⁰

The analysis here showed that incidence of detected bacterial infection, *where adequate sampling had taken place*, was 83% in the neutrophil driven endotype. The criteria for adequate sampling was at least four of the above domains for bacterial infection detection. This value, although much higher than in the other endotypes, was not statistically significant (OR=3.93, 95% CI 0.71-21.7; $p = 0.54$, Table 5.3). There was evidence of bacterial infection in the other endotypes (adaptive = 56% samples, endothelial leak = 57% samples), so secondary infection would not account for the features identified in the neutrophil driven endotype on its own.

To determine whether gene modules were associated with bacterial infection, Pearson's correlation coefficient was calculated between infection status and gene module eigenvalues. This was possible in 69 patients where sampling for bacterial infection was deemed to have been adequate according to the study criteria. Only the 'dark red' module was significantly associated with proven bacterial infection ($r = 0.48$, adjusted $p = 0.0004$). This module, consisting of 125 transcripts, was not significantly associated with a known process or pathway (Table 3.7).

The 'midnight blue' gene module, which was associated with worse clinical outcomes and enriched for the process 'antimicrobial humoral response', had a correlation coefficient equal to 0.1 (adjusted $p = 1$) for confirmed bacterial infection. The processes associated with this gene module could not be used to infer secondary bacterial infection.

5.3 HARP-2 endotypes

5.3.1 Clinical features of endotypes identified in the HARP-2 study

The HARP-2 study collected samples for biomarker analysis from 511 patients on the recruitment day of the trial. The results from the measurement of six biomarkers were available for this analysis. Hierarchical clustering with Ward linkage identified three clusters (Figure 3.12). Two of these clusters were associated with high levels of IL-6 and Ang-2, but were differentiated their relative levels of MMP-8 and sRAGE and were termed ‘MMP-8 dominant’ and ‘sRAGE dominant’ respectively. The third cluster showed depressed concentrations in five of the six biomarkers and was termed ‘hypo-inflammatory’.

There were no differences in baseline characteristics (demographics, ARDS aetiology) between patients in the three endotypes (Table 3.4). Patients with the MMP-8 and sRAGE dominant endotypes were associated with worse organ dysfunction with respect to serum creatinine and bilirubin levels, platelet counts and requirement of renal replacement therapy (RRT) when compared with patients in the hypo-inflammatory endotype (Figure 5.9 and Table 5.4). The sRAGE dominant endotype was associated with lower PaO₂-FiO₂ ratio ($p = 0.02$) and higher APACHE-II scores ($p = 0.005$). The MMP-8 endotype was associated with administration of higher vasopressor doses which was determined by a cardiovascular SOFA score greater than two (OR 2.2, 95% CI 1.4-3.4).

Patients with the hypo-inflammatory endotype had significantly lower C-reactive protein (CRP) levels than the other two endotypes which was consistent with its lower IL-6 levels (Figure 5.9). CRP, a circulating marker of inflammation, was not used to derive these endotypes. The MMP-8 dominant endotype had significantly higher CRP levels than the sRAGE dominant endotype ($p = 0.04$) which was surprising given their relative levels of IL-6 were similar. IL-6 stimulates release of CRP from the liver in the acute phase of the inflammatory response¹⁸⁷ and so a difference here was unexpected. This suggested that other, unmeasured mediators, were stimulating CRP release in these patients.

	MMP-8 dominant	sRAGE dominant	Hypo-inflammatory
N (recruitment day samples)	160	89	262
Sex = Male (%)	87 (54%)	50 (56%)	152 (58%)
Age years, mean (sd)	56 (16.6)	51.6 (15.6)	53.4 (16.5)
BMI, mean (sd)	26.6 (5.6)	27.2 (6.8)	27.6 (7.5)
Randomised to simvastatin	71 (44.4%)	44 (49.4%)	132 (50.4%)
ARDS Aetiology			
Pneumonia	82 (51.3%)	41 (46.1%)	157 (59.9%)
Sepsis	37 (23.1%)	21 (23.6%)	35 (13.4%)
Gastric aspiration	14 (8.8%)	8 (9.0%)	26 (9.9%)
Other	10 (6.3%)	8 (9.0%)	21 (8.0%)
Pancreatitis	10 (6.3%)	3 (3.4%)	3 (1.1%)
Thoracic trauma	6 (3.8%)	7 (7.9%)	16 (6.1%)
Non-thoracic trauma	1 (0.6%)	1 (1.1%)	4 (1.5%)
Clinical variables			
Creatinine (umol/L)	95 [68-155]	85 [62-130]	69 [54-107]
Bilirubin (umol/L)	15 [9-26.5]	15 [17-35]	9 [6-17.3]
Platelet count (x10 ⁹ /L)	163 (114)	163 (110)	220 (118)
C-reactive protein (mg/L)	226 [152-295]	179 [114-258]	141 [79-222]
Organ Dysfunction			
APACHE-II Score, median, [IQR]	19 [15-24]	21 [16-27]	17 [13-23]
PaO ₂ -FiO ₂ ratio, median [IQR]	18.5 [12.9-23.6]	17.6 [12.1-22.3]	19.3 [13.9-26.2]
		p = 0.02	Reference
Cardiovascular SOFA ≥ 3	124 (77.5%)	62 (69.7%)	160 (61.1%)
	OR 2.2 [1.4-3.4]		Reference
Renal replacement on Day 1	43 (26.9%)	20 (22.5%)	17 (6.5%)
	OR 5.3 [2.9-9.7]	OR 4.18 [2.08-8.41]	Reference
Any renal replacement therapy	68 (42.5%)	32 (36%)	36 (13.7%)
	OR 4.6 [2.9-7.4]	OR 3.5 [2.0-6.16]	Reference

Table 5.4 Characteristics of patients in each of the HARP-2 endotypes. The hypo-inflammatory endotype was associated with less organ dysfunction and better outcomes than the other two endotypes. ANOVA with Tukey's *post hoc* test was used to compare continuous variables. Categorical variables were compared using logistic regression. Length of stay was compared using a log-linear model.

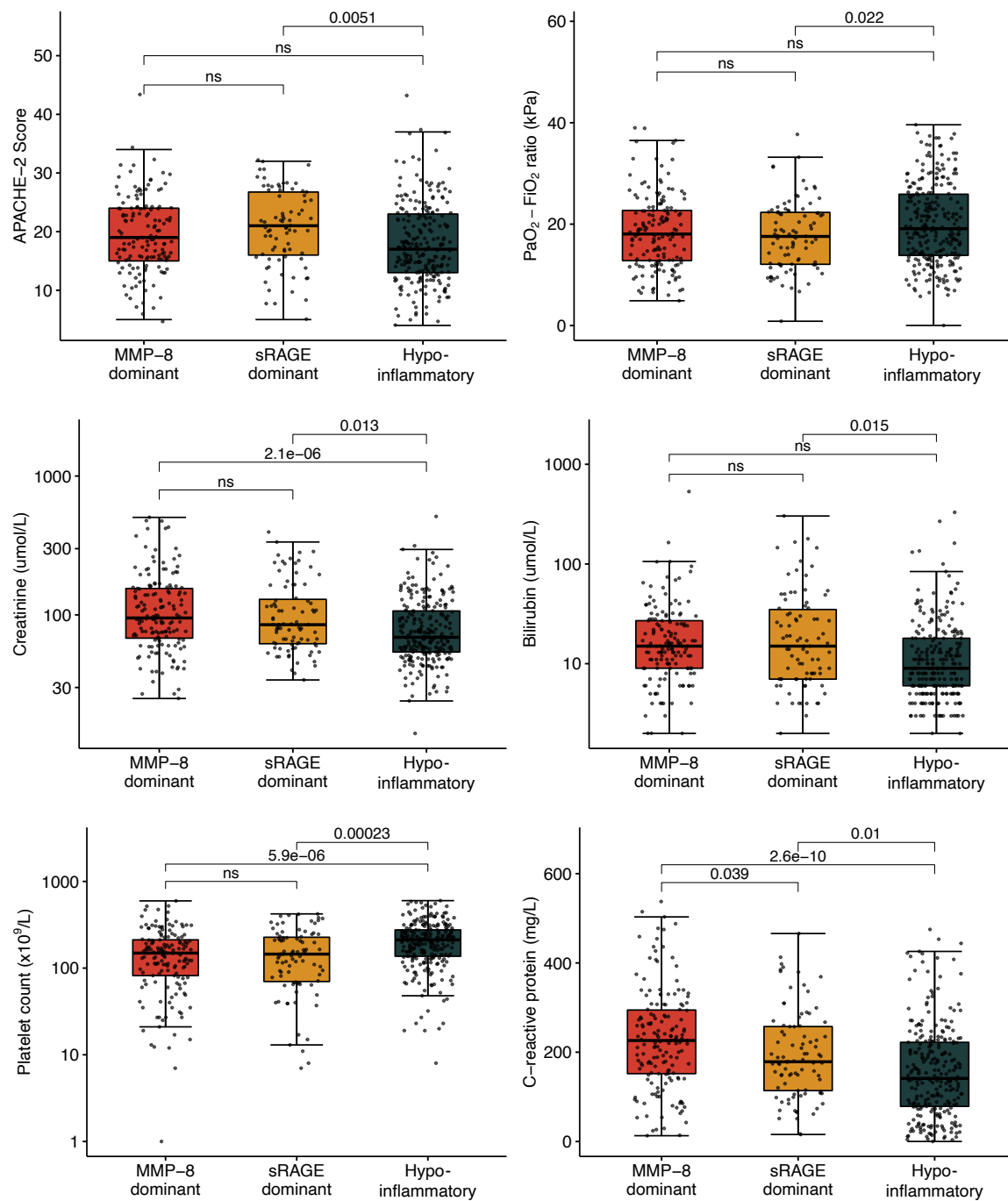


Fig. 5.9 Boxplots of clinical variables related to organ dysfunction in patients from each endotype identified in the HARP-2 study.

5.3.2 HARP-2 endotypes are associated with different outcomes and treatment response

Patients with the MMP-8 driven and sRAGE driven endotypes had worse outcomes than patients with the hypo-inflammatory endotype (Table 5.5). All measures of mortality, even with adjustment for organ dysfunction using the APACHE-II score and patient age, were worse in patients with the MMP-8 driven and sRAGE driven endotypes.

Figure 5.10 depicts the 28-day survival curves for each endotype. Panel **A** shows that patients with the MMP-8 driven and sRAGE driven endotypes followed a similar trajectory. Panel **B** shows the survival of patients who were randomised to the placebo arm of the study. The survival curves closely resembled the trajectories seen in panel **A**. Panel **C** shows that patients with the MMP-8 driven endotype had a treatment response to simvastatin. The survival curve of these patients resembled that of the hypo-inflammatory group treated with simvastatin. The 28-day survival of patients with the sRAGE driven and hypo-inflammatory endotypes was unchanged between the placebo and simvastatin arms, suggesting these patients did not respond to treatment.

The effect of simvastatin treatment on patients with each endotype was quantified, using Cox proportional hazards, in Table 5.5. The adjusted hazard ratio for patients with the MMP-8 driven endotype was equal to HR 0.35 (95% CI 0.18-0.71, $p = 0.003$), which implied a 65% risk reduction of death at 28 days. Recruitment of patients with the other two endotypes, who were treatment unresponsive, therefore masked the treatment response shown by these patients and led to negative trial result. If treatment effects were uniform in all patients with this endotype, then the expected 28-day mortality in the MMP-8 driven endotype might have reduced from 31.9% to 11.2%. Caution must always be advised when conducting *post hoc* sub-group analysis of randomised controlled trials, but even with this caveat, these results are compelling.

	MMP-8 Dominant	sRAGE dominant	Hypo-inflammatory
N	160	89	262
Randomised to simvastatin	71 (44.4%)	44 (49.4%)	132 (50.4%)
Patient outcomes			
Length of ICU stay [§] median days, [IQR]	13 [8-21] p <0.001	9 [6-16.8]	9 [5-16] Reference
28 day mortality	51 (31.9%)	30 (33.7%)	43 (16.4%)
<i>Unadjusted</i>	OR 2.38 [1.5-3.8]	OR 2.6 [1.5-4.48]	Reference
<i>Adjusted</i>	OR 2.32 [1.38-3.9]	OR 2.66 [1.44-4.92]	Reference
ICU mortality	44 (27.5%)	34 (38.2%)	41 (15.6%)
<i>Unadjusted</i>	OR 2.04 [1.26-3.31]	OR .33 [1.93-5.7]	Reference
<i>Adjusted</i>	OR 1.93 [1.13-3.28]	OR 3.39 [1.85-6.22]	Reference
Hospital mortality	56 (35%)	35 (39.3%)	56 (21.3%)
<i>Unadjusted</i>	OR 1.98 [1.28-3.07]	OR 2.38 [1.42-4]	Reference
<i>Adjusted</i>	OR 1.70 [1.10-2.93]	OR 2.41 [1.34-4.33]	Reference
Treatment response to simvastatin			
Unadjusted	HR 0.43 [0.24-0.79] p = 0.007	HR 1.65 [0.80-3.42] p = 0.18	HR 0.90 [0.49-1.64] p = 0.72
Adjusted for APACHE-II and Age	HR 0.35 [0.18-0.71] p = 0.003	HR 1.42 [0.66-3.05] p = 0.37	HR 0.84 [0.43-1.64] p = 0.61

Table 5.5 Outcomes of patients in each of the HARP-2 endotypes. The hypo-inflammatory endotype was associated with better patient outcomes across all measures. Treatment response to simvastatin refers to the 28 day mortality of patients in each endotype who were randomised to the simvastatin arm of the trial, compared with those, from the same endotype who were randomised to placebo. Adjusted models included APACHE-II score and patient age as covariates for fitting logistic regression (OR) or Cox proportional hazards models (HR). The MMP-8 dominant endotype showed a significant treatment response to simvastatin. [§]Length of stay in survivors.

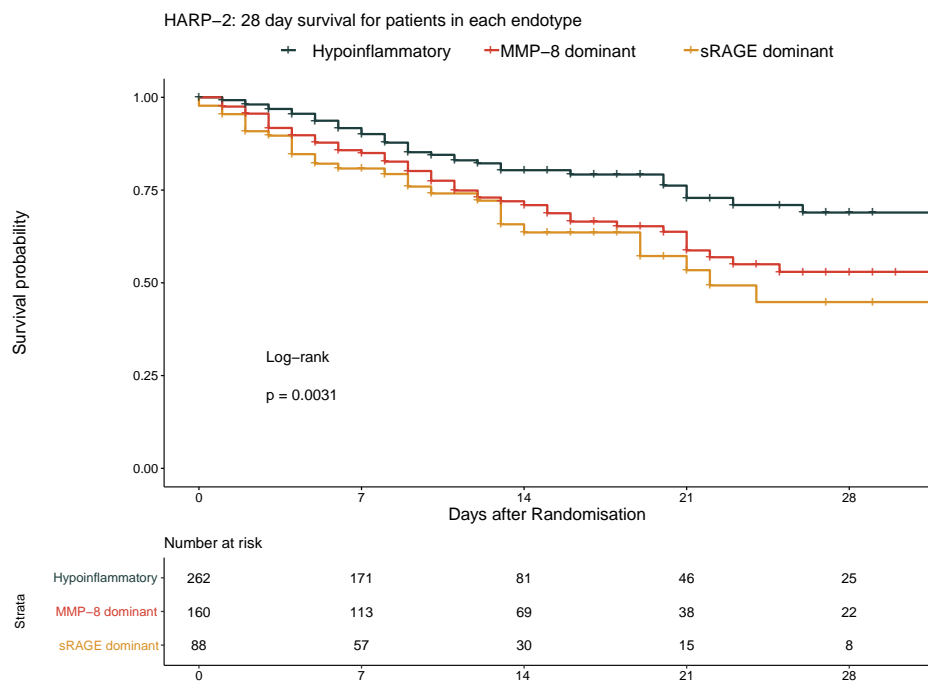
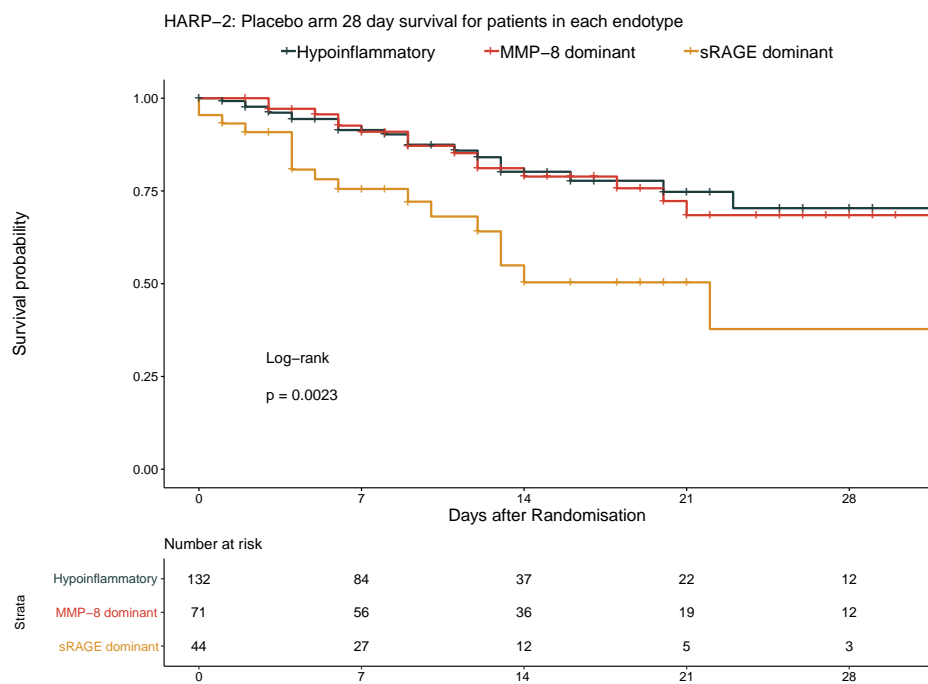
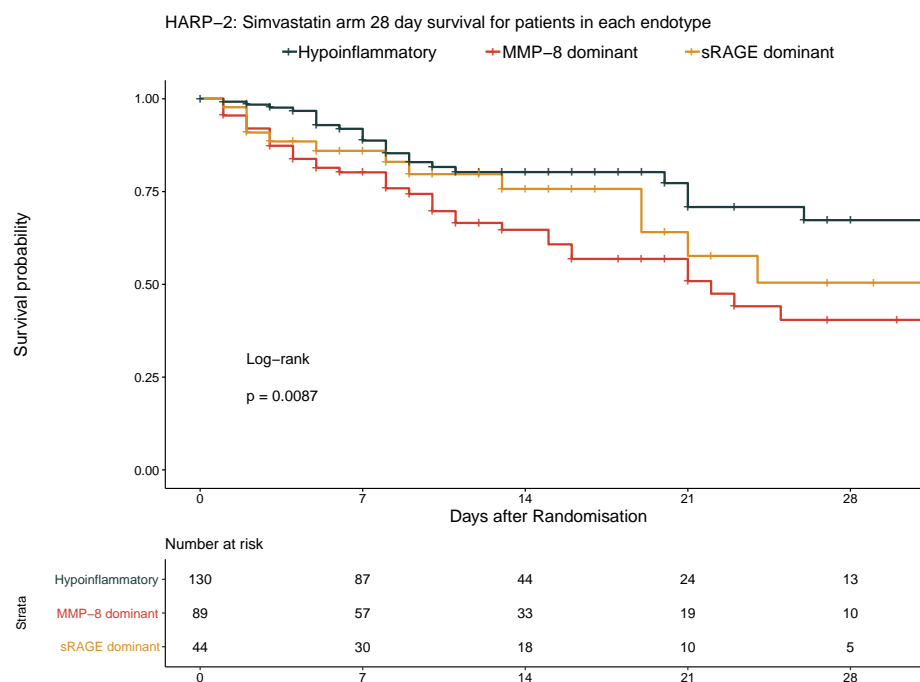


Fig. 5.10 (a)



5.10 (b)



5.10 (c) Kaplan-Meier analysis of patients in each HARP-2 endotype. **a** shows that patients with the hypo-inflammatory endotype had significantly better 28 day survival compared with patients with the other two endotypes. The same was true for the patients randomised to the placebo arm (**b**). **c** shows that the patients with the MMP-8 dominant endotype had a significant therapeutic response to simvastatin, and their outcomes were comparable to patients with the hypo-inflammatory endotype.

5.4 Summary and discussion of endotype characterisation

Isolated biological processes, determined at a gene expression level, are generally considered far removed from the organ dysfunction and loss of homeostasis required to cause abnormal values in blood tests and organ failure requiring support on intensive care. Measurement of gene expression and blood protein biomarkers provides a global assessment of circulating immune cell function. Immune dysfunction plays a crucial role in sepsis, ARDS and other critical care syndromes, and these mechanisms may contribute to the pathology and organ dysfunction patients suffer. The analysis in Chapters 3.2 and 4 described some of the processes that characterised endotypes. This chapter outlined the clinical features of endotypes and determined whether these translated into clinical phenotypes and affected patient outcomes.

5.4.1 GAinS

The only statistically significant difference between patients in each of the GAinS endotypes was a lower PaO₂-FiO₂ ratio for those with the hyper-inflammatory endotype compared with those with intermediate endotype ($p = 0.007$, Figure 5.1). Patients with the hyper-inflammatory endotype had lower bicarbonate, platelet counts and higher creatinine than those with the hypo-inflammatory endotype. These were not significant after adjustment for multiple comparisons. The trends for these three variables between these two endotypes, although not significant, were consistent with the polarised immune responses in these patients. The hypo-inflammatory endotype consisted of only 27 patients (19%) and so a failure to detect a difference between the hypo-inflammatory and the hyper-inflammatory endotype may be due to under-powered statistical comparisons.

Patients with the hypo-inflammatory endotype were more likely to have received steroid therapy (OR = 3.5, 95% CI 1.3-9.8, Table 5.1) than those with the intermediate endotype. This was consistent with the globally suppressed cytokine response in these patients. Further corroboration of the role of steroids in this endotype was provided by two important transcripts related to glucocorticoid receptors being present in the gene module that differentiated these patients from the hyper-inflammatory endotype: *FKBP5*, *SGK1*. The gene *FKBP5* contains steroid responsive elements in its promoter regions and its expression is rapidly increased following glucocorticoid receptor activation.¹⁸⁸

The clinical data provided from the GAinS study did not discriminate between patients who were receiving chronic steroid treatment from those who received supplemental steroid therapy as part of their critical care management. Critical care patients with refractory

shock are often administered intravenous hydrocortisone in the acute setting to improve the efficacy of exogenous catecholamines for cardiovascular support. If patients with the hyper-inflammatory endotype received steroid therapy in this context, these patients might not have developed an immunosuppressed phenotype. Chronic steroid treatment might have accounted for the biomarker profile in the hypo-inflammatory endotypes. However, ambiguous recording of steroid treatment in patients recruited to the GAINs study may have masked the relative effects of acute and chronic steroid administration.

5.4.2 MOSAIC

Patients with the neutrophil driven endotype in the MOSAIC study were more likely to have: lower platelet counts, higher creatinine and bilirubin. These patients were also more likely to require mechanical ventilation, vasopressor support and had worse outcomes than those with the adaptive endotype (Tables 5.2 and 5.3). The low albumin levels in the endothelial leak endotype was a feature that distinguished these patients. There had been very little evidence to suggest that there were any differences between patients with the neutrophil driven and endothelial leak endotypes based on the distribution of clinical variables and differential gene expression. However, the Kaplan-Meier analysis showed the 30-day survival of the patients with the endothelial leak endotype to be significantly higher than the hyper-inflammatory endotype, and similar to those with the adaptive endotype. This was unexpected because patients with the endothelial leak endotype were significantly more likely to require ventilation (OR = 3.68, 95% CI 1.44-9.8) and require cardiovascular support (OR = 6.11, 95% CI 2.2-16.9) than patients with the the adaptive endotype. Confidence intervals of these statistics were wide because of the small number of patients in each endotype.

The patients in the endothelial leak endotype might represent the ‘quick’ turnaround subset of critical care patients who only require brief periods of organ support and do not develop complications that prolong their stay on ICU. Endothelial leak might be offset by positive end-expiratory pressure (PEEP) from invasive ventilation if this was the process that caused their need for admission to critical care. In contrast, patients with neutrophil driven endotype may represent the subset with the greater immune dysregulation who develop multi-organ dysfunction and have worse outcomes. Whether the immune dysregulation caused or was the result of worse multi-organ dysfunction cannot be determined without repeated temporal sampling.

Another possible explanation for the difference between endotypes might be secondary bacterial infection in patients with the neutrophil driven endotype. Diagnosis of bacterial infection is challenging in critical care patients as it is difficult to sample the affected organs

(usually the lung) and fastidious organisms may not grow in cultured samples because patients are often receiving antibiotics. Patients with the neutrophil driven endotype had higher levels of procalcitonin (PCT), a biomarker sensitive for bacterial infection. In the context of severe pandemic H1N1 influenza PCT can be a poor discriminator for determining bacterial super-infection in these patients, who may have PCT concentrations as high as 10 ng/mL without detectable secondary infection.¹⁸⁹ The manufacturers of the PCT assay consider >0.5 ng/mL as the threshold for clinical suspicion of bacterial infection.

Although the proportion of patients with bacterial infection was higher in the neutrophil driven endotype, this was not statistically significant. Additionally, there was evidence of bacterial infection in the other two endotypes. Bacterial infection, therefore, likely contributed to the immune responses observed in the neutrophil driven endotype but did not fully account for the outcomes and features associated with these patients.

This analysis successfully demonstrated stability and persistence of endotypes in these patients at 48 hours. There is an opportunity to replicate these endotypes prospectively in future studies, which might offer the opportunity to stratify patients to interventions in clinical trials.

5.4.3 HARP-2

The endotypes identified in the HARP-2 study were demonstrated using biomarkers measured at the time of recruitment to the study. Similar patterns emerged in this study as with the other two: there were hyper-inflammatory and hypo-inflammatory endotypes, and the hypo-inflammatory endotype was associated with less organ dysfunction and better outcomes (Table 5.4 and 5.5).

Due to the large number of patients with biomarker measurements in the HARP-2 study these results were based on larger groups and so the confidence intervals of the statistical estimates are narrower, in contrast with the comparisons made between endotypes in other two studies.

Derivation of endotypes in the HARP-2 study used biomarkers that were not measured in the MOSAIC or GAIN studies (MMP-8, Ang-2, sRAGE, SP-D). These additional biomarkers differentiated the patients with a hyper-inflammatory subtype into MMP-8 dominant and sRAGE dominant endotypes. These two endotypes were similar with respect to organ dysfunction measures and outcomes (Tables 5.4 and 5.5, Figure 5.9). The primary difference between these two endotypes was their response to treatment with simvastatin. The MMP-8

driven endotype showed a treatment benefit with simvastatin therapy that aligned their 28 day survival curve with the survival curve of patients with the hypo-inflammatory endotype.

The MMP-8 driven and sRAGE driven endotypes were both associated with high levels of IL-6, sTNFR-1s and Ang-2. Ang-2 is a marker of endothelial injury. The proposed mechanism by which simvastatin was expected to benefit patients with ARDS was to modulate endothelial integrity and reduce inflammation.¹¹¹ A possible explanation for why some patients responded to simvastatin might be related to the high levels of sRAGE in the sRAGE driven dominant endotype.

sRAGE is the soluble form of RAGE, a pattern recognition receptor which has a similar structure to immunoglobulin. Levels of sRAGE are raised in a number of inflammatory and non-inflammatory conditions including arthritis, Alzheimer's disease, sepsis, COPD and obesity.¹⁹⁰ It is however principally found in lung tissue where it is expressed by type 1 alveolar cells. The high levels of sRAGE in patient with the sRAGE driven may reflect worse disruption of the endothelium that is compromised to the extent that large amounts of sRAGE enter the circulation. It could be hypothesised that this degree of endothelial disruption was no longer amenable to modulation by simvastatin. sRAGE might therefore be used as a biomarker of predicting treatment failure with simvastatin in the context of this study. Similarly, if patients with the hypo-inflammatory endotype could be identified at randomisation, these patients might be excluded from immunosuppressive therapies. These patients have a relatively better outcome at 28 days and so are less likely to derive benefit from interventions but may suffer adverse events.

Restriction of analysis of the HARP-2 results to recruitment day samples permits the findings in this analysis to be applied to future randomised controlled trial design for intensive care patients. This would require measurement of protein biomarkers as part of a predictive enrichment strategy at the time of recruitment, prior to randomisation.

CHAPTER 6

General Discussion and Conclusions

6.1 Summary of endotypes

Figure 6.1 provides a summary of the clinical characteristics and outcomes of patients in each endotype, alongside the biological mechanisms that they are associated with.

Three endotypes were identified in the GAINs study. These were termed: hyper-inflammatory, intermediate and hypo-inflammatory. The hyper-inflammatory endotype was associated with dysregulated cytokine release and gene networks implicated in HLH. The hypo-inflammatory endotype was associated with global suppression of cytokine release and treatment with glucocorticoids may explain the low cytokine concentrations in these patients.

Three endotypes were identified in the MOSAIC study. These were termed: endothelial leak, adaptive and neutrophil driven. The endothelial leak endotype was associated with SLIT-ROBO signalling, low albumin levels and a favourable 30-day survival despite the need for organ support. The adaptive endotype was associated with lymphocyte and IFN- α 2a mediated immune responses and a lower requirement for organ support. The neutrophil driven endotype was associated with multi-organ dysfunction and expression of genes associated with neutrophil activation and degranulation. These endotypes were not fully explained by the relative incidence of confirmed secondary bacterial infection.

Three endotypes were identified in the HARP-2 study: these were termed hypo-inflammatory, MMP-8 driven and sRAGE driven. The hypo-inflammatory endotype was associated with lower serum concentrations of IL-6 and sTNFR-1 and lower 28-day mortality. The MMP-8 driven endotype was associated with raised IL-6 and MMP-8 and demonstrated a favourable treatment response to simvastatin. The sRAGE driven endotype was associated with raised IL-6 and sRAGE concentrations. Patients with the MMP-8 driven and sRAGE driven endotypes

has similar outcomes, but the sRAGE driven endotype did not demonstrate any treatment response to simvastatin.

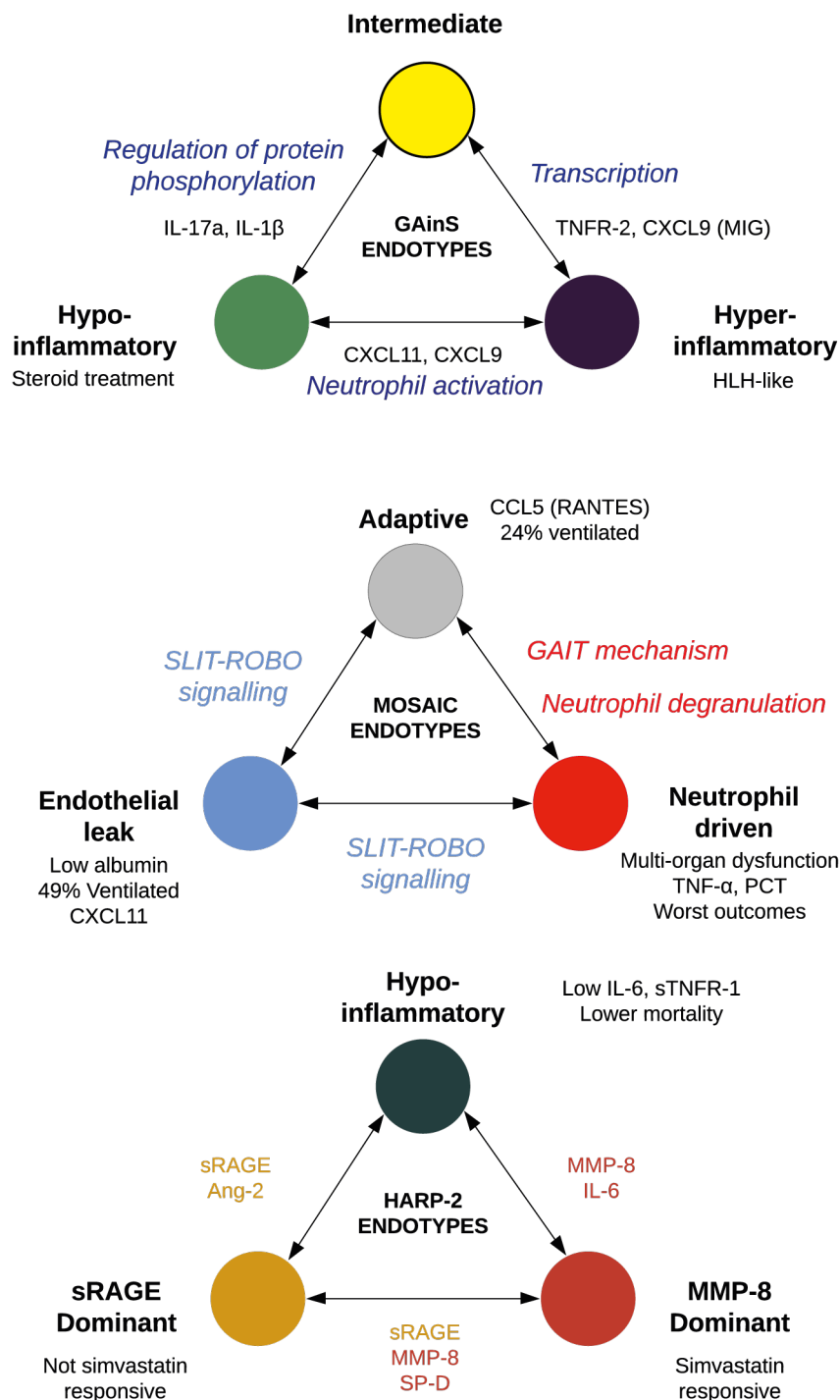


Fig. 6.1 Final endotype model from each contributing study

There are several common themes here between all three studies:

- A three endotype model captures some of the heterogeneity in critically unwell patients.
- Hypo-inflammatory subsets of critical illness are apparent in all three studies. These patients generally had less organ dysfunction and better outcomes.
- Hyper-inflammatory states can be sub-classified into plausible endotypes that have different outcome profiles and responses to treatment on intensive care.
- Patients with critical illness syndromes can be separated into subgroups that are defined by different biological process, by using unbiased approaches that do not depend on clinical variables.

The favourable survival profiles of the relatively hypo-inflammatory HARP-2 and MOSAIC endotypes is in contrast to the findings of other groups. For example, the SRS-1 phenotype, derived from the same GAINs transcriptomic data, was associated with worse patient outcomes. The authors determined this phenotype was related to T-cell exhaustion, depressed immunity and described it as a hypo-inflammatory state.⁹⁸

The findings of the MARS consortium sepsis study identified a favourable endotype associated with adaptive immunity and interferon signalling (MARS-3). This endotype bore some resemblance to the SRS-2 endotype found by the GAINs consortium.⁹⁶ These themes fit with the MOSAIC adaptive endotype described here.

Patients with the hyper-inflammatory (Calfée et al., (2014)) or the hyper-reactive (Bos et al., (2017)) ARDS endotypes had worse outcomes in their respective studies.^{61,82} Calfée et al.'s hyper-inflammatory endotype was responsive to treatment with simvastatin¹⁰⁶ Both of these are consistent with the neutrophil driven and hyper-inflammatory endotypes described in this thesis.

One strength of this analysis is that even with relatively few biomarkers, novel endotypes which have implications for patient management can be described. The hyper-inflammatory endotypes observed may benefit from immunosuppressive therapy, whilst the hypo-inflammatory endotypes might be excluded from immunosuppressant interventions. Not least because their background survival is better than the other endotypes, but also because they might be less likely to benefit and may suffer adverse reactions.

Inclusion of patients belonging to a non-responsive endotypes in a randomised controlled trial may negate the positive effects in patients with treatment responsive endotypes, leading to negative or equivocal trial outcomes. Critical care patients frequently suffer from the adverse

effects of iatrogenic interventions: nosocomial infections, delirium, ventilator-induced lung injury to name but a few. Recognition of which patients might come to additional harm from therapeutic medications is as important as the identification of treatments which improve outcomes.

In this analysis, attempts have been made to fully characterise the mechanisms underlying each identified endotype, using a combination of differential gene expression and linear discriminant analysis. Whilst other studies use the hyper-inflammatory and hypo-inflammatory labels, I have shown that more nuance may need to be applied when using these terms. For example, the ‘grey’ cluster from MOSAIC study might have been described as hypo-inflammatory due to its relatively low concentrations of IL-6 and TNF- α . Differential gene expression and LDA demonstrated that patients with the adaptive endotype were associated with lymphocyte activation and adaptive immunity. Similarly, a three endotype model of HARP-2 showed that there were two hyper-inflammatory endotypes which had different responses to treatment with simvastatin.

Another strength of this analysis is temporal stability of endotypes identified in the MOSAIC study. Stability implies that these immune states were present for at least two days. The immune profiles observed are in the context of a heterogeneous population of influenza patients who were recruited at different times during the course of their acute illness. The immune profiles must therefore be persistent and lends to support to the hypothesis that immune responses in critical illness are stereotyped.

This finding offers the opportunity to prospectively identify patients with these endotypes at the time of recruitment to an interventional study. Predictive enrichment to an RCT by stratification in this manner has the potential to maximise the utility of expensive trials and novel treatments. Delucchi et al. (2018) have shown that latent class-based endotypes, identified in patients with ARDS, were stable over three days.¹⁰⁷ Kitsios et al. (2019) have demonstrated endotype stability, using latent class analysis of a ten biomarker panel, for as long as two weeks.¹⁹¹

The underlying mechanisms demonstrated here might be amenable to experimental verification using laboratory-based techniques. The endothelial leak endotype might be observed with *in-vitro* leak assays of pulmonary microvascular endothelial cells. Serum from patients with different MOSAIC endotypes and healthy controls could be tested to see if they induce endothelial leak and whether this effect could be abrogated by addition of Slit-2. These experiments would be similar to those conducted by London et al. (2010) and Weng et al. (2019).^{169,171} The HLH-like endotype identified in the GAinS study could be confirmed by measuring soluble IL-2 receptor (CD25), triglycerides and ferritin. Measurement of

caeruloplasmin in samples from the adaptive and neutrophil driven MOSAIC endotypes might demonstrate whether the GAIT system is functioning in the context of raised IFN- γ levels.

6.2 Limitations

The analysis I have undertaken is subject to several limitations that might preclude its external validity.

There were no differentially expressed genes between patients with or without ARDS, nor between clusters in the GAINs microarray results. This was surprising considering how polarised the immune responses in these patients were. It is likely that there was additional heterogeneity amongst patients with the protein biomarker-based endotypes that was not fully captured by measurement of cytokines and chemokines. Other possible explanations might include use of batch effect correction methods that may have suppressed probe variances. The methods used here were no different to standard methods used in the literature and by the GAINs research group in their publications.

Additional steps that might have been taken to improve the detection of differences in gene expression include: restricting the number of probes used and removal of outlier samples. Review of the MDS plots of the gene expression profiles identified approximately 15 samples that might be considered outliers (Figure 3.15). This relatively small number of outliers is unlikely to have suppressed genuine biological signals considering the number of samples in each endotype were relatively large.

Another limitation is the assumption upon which co-expression network analysis relies: that the network structure has scale-free properties. Although some biological networks may be scale-free, it is increasingly recognised that this may not be the case.¹⁹² WGCNA is however still widely used by researchers in the field of transcriptomic analysis.^{127,182,193} Alternative methods of network analysis tend to be supervised, based on previously described annotations and protein-proteins interactions, or are still maturing in their development.¹⁹⁴ These alternative methods may also have similar flaws due to the assumptions upon which they are based.¹⁹⁵

Linear discriminant analysis models demonstrated adequate performance in this analysis because the protein biomarker and module eigengene data happened to be orthogonal when combined and linearly separable. There is no guarantee this would be the case in other studies. The variables, using which the LDA models were fitted, were log-transformed and centre

scaled with zero mean. This zero mean was therefore derived from the study population, not control samples.

This was not problematic for this analysis as the goal was to characterise the heterogeneity of critically unwell patients in these three studies. However, if these results were to be translated into use for in a clinical trial, the population means of each cytokine or biomarker would have to be determined in the study population before stratification. The methods used for the measurement of protein biomarkers are not currently licensed for use in patient diagnosis. There are no standardised reporting methods for these biomarker assays. Standardisation allows for calibration between laboratories using different quantification methods. For these reasons, thresholds for concentrations of protein biomarkers that might predict endotype membership were not calculated in this analysis.

The latent class analysis method used by the Calfee research group also uses scaled and ordinal transformation of continuous variables to derive endotypes.⁶¹ Validation of these unsupervised learning approaches will therefore require parsimony across multiple studies.

The fitted LDA models performed well in this analysis because classes were easily separable using a linear transformation. Noisier data would have led to poor model performance, and different methods might have been required to distinguish clusters in a more robust manner. The application of LDA here offered transparency as to its method and straightforward interpretation with regards to its outputs. The goal of this thesis was the discovery of the important endotype features, not prospective prediction with repeated cycles of model tuning.

Hierarchical clustering with Ward linkage of protein biomarker concentrations found spherical clusters and divided these data into subsets with relatively extreme values. This would also explain why the LDA method was effective in discriminating these clusters. Complex biological data does not always fit neatly into spherical groups, and so edge cases will be misclassified by this method. On the other hand, alternative clustering methods (DBScanⁱ, OPTICSⁱⁱ) which perform well with multi-dimensional data are optimised to identify genuinely separate clustered groupings.^{103,196} The PCA representation of the cytokine data from all three studies showed that the data points formed a fairly homogeneous sphere in the first few visualised dimensions. There was little apparent separation into groups that might be identified as distinct clusters by the DBScan or OPTICS methods.

ⁱDensity-based spatial clustering of applications with noise

ⁱⁱOrdering points to identify the clustering structure

The Ward-linkage clustering method I used worked well by exposing the heterogeneity in critically unwell patients' immune profiles. This method of clustering is subject to instability, as small perturbations in the data or removal of features may produce new clustering assignments that are unrelated to the original clusters. To avoid this, large sample sizes and many variables are required to ensure clusters are robust. Bootstrapped resampling and measurements of the adjusted Rank index were used to quantify instability and ensure clusters were robust.

An alternative approach to hierarchical clustering might have involved the projection into a different coordinate system, using an embedding process. T-distributed stochastic neighbour embedding (tSNE) or uniform manifold approximation and projection (UMAP) are examples of alternative embedding methods. Clustering could then be applied to the new embedded space. This approach is often used in the analysis of single-cell RNA sequencing and other complex 'omics data. Embedding-based approaches were avoided in this thesis because the embedding process cause the features of the data to become difficult to interpret once transformed and projected in this way.

Enrichment analysis of gene modules from the MOSAIC and GAINs samples tended to favour processes and pathways associated with neutrophil function and activation. This was expected given that the patients from which these samples were taken had either sepsis or severe influenza infection. However, these enrichment results may have simply been a function of the relatively greater abundance of neutrophils circulating in these patients. Therefore, this enrichment strategy might have been less likely to favour cells that are poorly represented in the whole blood transcriptome.

The results from the MOSAIC study may have demonstrated this: patients in the 'red' (neutrophil driven) MOSAIC cluster had high levels of IL-15. This cytokine is associated with NK cell activation and recruitment in influenza infection.¹⁴⁴ Any influence that NK cells may have had on the underlying mechanisms in these patients was over-shadowed by the relative abundance of neutrophil related processes.

Adjustment of expression levels for neutrophil counts using a linear model might control expression levels for the relative of individual cell types. This approach is not feasible if applying WGCNA, unless all probe levels were adjusted for neutrophil count, in which case the low abundance signatures from minor cell types will also be depressed. The ideal approach would be to use concurrent immunophenotyping to adjust for all cell types. Single-cell RNA sequencing may address some of these problems in the future as this technology evolves.

Prospective replication of these results may not be possible as the microarrays chip versions used to quantify gene expression in the MOSAIC and GAinS studies are no longer available. Research exploring gene expression in human studies now uses bulk RNA sequencing methods (RNAseq), which is a different technology that require separate analytical tools. A broader problem with RNA-based technologies, in general, is the abundance of transcripts may not be reflective of protein levels. Post-transcriptional, translational and other regulatory steps in protein synthesis may affect protein production and function. Sequencing methods cannot account for these additional biological processes. Integration of proteomic, transcriptomic and clinical variables will be necessary to fully describe and characterise the processes which define endotypes.

6.3 Validity of this approach in future studies

The goal of this thesis was to discover the most important biological mechanisms in critically unwell patients. Disease labelling and stratification according to clinical features was avoided at early stages of the analysis. Critical to the approach was not the method of clustering or network analysis, but the recognition that labels and outcomes of patients are subject to biases when used in supervised learning. This unbiased approach could be applied to other heterogeneous diseases.

The results from the HARP-2 study demonstrated a treatment responsive group could be replicated in other studies. Extended cytokine analysis is being conducted on the HARP-2 samples in order to improve the characterisation of each of the identified endotypes here and determine whether similar endotypes to those found in the MOSAIC and GAinS studies can be replicated in the HARP-2 study.

Building on this work requires prospective validation of these mechanisms before predictive models can be fitted to stratify patients accurately. The clusters identified here were defined, initially, by their different immune profiles after measurement of 25 or more mediators. Although an enriched subset of cytokines might predict endotypes better, the differences between them are subtle. There were only a relative handful of differentially expressed genes between clusters with significant differences in cytokine levels and survival profiles. Point of care stratification of critical care patient using protein biomarkers alone may befall problems with heterogeneity and poor reproducibility unless the precise mechanisms are better understood.

6.4 Summary and conclusion

I have demonstrated that there are distinct biological profiles in critically unwell patients and that by using an integrated approach, they can be defined by their underlying mechanisms. I have been able to make high-level insights into the different pathological processes that are likely to be contributing to critical illness syndromes in these patients. The mechanisms described are plausible, as they are supported by clinical or biomarker data and are amenable to experimental verification or prospective validation in patients. These endotypes are clinically relevant as they are associated with the differing levels of organ dysfunction, patient outcomes and treatment response.

In this thesis, I analysed transcriptomic data with network methods and used a novel approach to integrate features from the down-sampled, enriched network of highly connected genes with protein biomarkers. Therefore, unsupervised learning approaches may help to disentangle the heterogeneity of critically unwell patients and characterise them based on biological features. If verified in future studies, the insights gained might enable future clinical trials to be prospectively enriched with these endotypes and allow for treatment stratification.

References

- [1] B. Taylor Thompson, Rachel C. Chambers, and Kathleen D. Liu. Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 377(6):562–572, 2017. doi: doi.org/10.1056/NEJMra1608077.
- [2] David G. Ashbaugh, D. Boyd Bigelow, Thomas L. Petty, and Bernard E. Levine. Acute Respiratory Distress in Adults. *The Lancet*, 290(7511):319–323, aug 1967. doi: doi.org/10.1016/S0140-6736(67)90168-7.
- [3] G. R. Bernard, A. Artigas, K. L. Brigham, J. Carlet, K. Falke, L. Hudson, M. Lamy, J. R. Legall, A. Morris, and R. Spragg. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am. J. Respir. Crit. Care Med.*, 149(3):818–824, 1994. doi: doi.org/10.1164/ajrccm.149.3.7509706.
- [4] V. Marco Ranieri, Gordon D. Rubenfeld, B. Taylor Thompson, Niall D. Ferguson, Ellen Caldwell, Eddy Fan, Luigi Camporota, and Arthur S. Slutsky. Acute respiratory distress syndrome: The Berlin definition. *JAMA*, 307(23):2526–2533, 2012. doi: doi.org/10.1001/jama.2012.5669.
- [5] J. F. Murray, M. A. Matthay, J. M. Luce, and M. R. Flick. An expanded definition of the adult respiratory distress syndrome. *Am. Rev. Respir. Dis.*, 138(3):720–723, 1988. doi: doi.org/10.1164/ajrccm/138.3.720.
- [6] G. D. Rubenfeld and M. S. Herridge. Epidemiology and outcomes of acute lung injury. *Chest*, 131(2):554–562, 2007. doi: 10.1378/chest.06-1976.
- [7] Giacomo Bellani, John G. Laffey, Tàì Pham, Eddy Fan, Laurent Brochard, Andres Esteban, Luciano Gattinoni, Frank M.P. Van Haren, Anders Larsson, Daniel F. McAuley, Marco Ranieri, Gordon Rubenfeld, B. Taylor Thompson, Hermann Wrigge, Arthur S.

- Slutsky, and Antonio Pesenti. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*, 315(8):788–800, 2016. doi: doi.org/10.1001/jama.2016.0291.
- [8] R. R. Thiagarajan, R. P. Barbaro, P. T. Rycus, D. M. McMullan, S. A. Conrad, J. D. Fortenberry, and M. L. Paden. Extracorporeal Life Support Organization Registry International Report 2016. *ASAIO J.*, 63(1):60–67, 2017. doi: doi.org/10.1097/MAT.0000000000000475.
- [9] Margaret S. Herridge, Angela M Cheung, Catherine M Tansey, Andrea Matte-martyn, Natalia Diaz-granados, Fatma Al-saidi, Andrew B Cooper, Cameron B Guest, C David Mazer, Sangeeta Mehta, Thomas E Stewart, Aiala Barr, Deborah Cook, and Arthur S Slutsky. One-Year Outcomes in Survivors of the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 348(8):683–93, 2003. doi: doi.org/10.1056/NEJMoa1207363.
- [10] A. M. Cheung, C. M. Tansey, G. Tomlinson, N. Diaz-Granados, A. Matte, A. Barr, S. Mehta, C. D. Mazer, C. B. Guest, T. E. Stewart, F. Al-Saidi, A. B. Cooper, D. Cook, A. S. Slutsky, and M. S. Herridge. Two-year outcomes, health care use, and costs of survivors of acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.*, 174(5):538–544, 2006. doi: doi.org/10.1136/thoraxjnl-2017-210217.
- [11] Margaret S. Herridge, Marc Moss, Catherine L. Hough, Ramona O. Hopkins, Todd W. Rice, O. Joseph Bienvenu, and Elie Azoulay. Recovery and outcomes after the acute respiratory distress syndrome (ARDS) in patients and their family caregivers. *Intensive Care Medicine*, 42(5):725–738, 2016. doi: doi.org/10.1007/s00134-016-4321-8.
- [12] D. W. Dowdy, M. P. Eid, C. R. Dennison, P. A. Mendez-Tellez, M. S. Herridge, E. Guallar, P. J. Pronovost, and D. M. Needham. Quality of life after acute respiratory distress syndrome: a meta-analysis. *Intensive Care Med*, 32(8):1115–1124, 2006. doi: doi.org/10.1007/s00134-006-0217-3.
- [13] Margaret S. Herridge, M Tansey, Catherine, Andrea Matte-martyn, George Tomlinson, Natalia Diaz-granados, Andrew B Cooper, Cameron B Guest, C David Mazer, Sangeeta Mehta, Thomas E Stewart, paul Kudlow, Deborah Cook, Arthur Slutsky, Angels Cheung, Deborah Cook, Arthur S Slutsky, Canadian Critical, and Care Trials. Functional Disability 5 Years after Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 364(14):1293–304, 2011. doi: doi.org/10.1056/NEJMoa1207363.

- [14] P. Johnson, W. Chaboyer, M. Foster, and R. van der Vooren. Caregivers of ICU patients discharged home: what burden do they face? *Intensive Crit Care Nurs*, 17(4):219–227, 2001. doi: doi.org/10.1054/icc.2001.1577.
- [15] B. B. Kamdar, K. A. Sepulveda, A. Chong, R. K. Lord, V. D. Dinglas, P. A. Mendez-Tellez, C. Shanholtz, E. Colantuoni, T. M. von Wachter, P. J. Pronovost, and D. M. Needham. Return to work and lost earnings after acute respiratory distress syndrome: a 5-year prospective, longitudinal study of long-term survivors. *Thorax*, 73(2):125–133, 2018. doi: doi.org/10.1136/thoraxjnl-2017-210217.
- [16] Marcelo B.P. Amato, Maureen O. Meade, Arthur S. Slutsky, Laurent Brochard, Eduardo L.V. Costa, David A. Schoenfeld, Thomas E. Stewart, Matthias Briel, Daniel Talmor, Alain Mercat, Jean-Christophe M. Richard, Carlos R.R. Carvalho, and Roy G. Brower. Driving pressure and survival in the acute respiratory distress syndrome. *New England Journal of Medicine*, 372(8):747–755, 2015. doi: doi.org/10.1056/NEJMSa1410639.
- [17] Acute Respiratory Distress Syndrome Network. Ventilation With Lower Tidal Volumes As Compared With Traditional Tidal Volumes for Acute Lung Injury and the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 342(18):1301–1308, 2000. doi: doi.org/10.1056/NEJM200005043421801.
- [18] M G Walker, L.-J. Yao, E K Patterson, M G Joseph, G Cepinskas, R A W Veldhuizen, J F Lewis, and C M Yamashita. The Effect of Tidal Volume on Systemic Inflammation in Acid-Induced Lung Injury. *Respiration*, 81(4):333–342, 2011. doi: doi.org/10.1159/000323609.
- [19] Pablo Cardinal-Fernandez, Jose A. Lorente, Aida Ballen-Barragan, and Gustavo Matute-Bello. Acute respiratory distress syndrome and diffuse alveolar damage new insights on a complex relationship. *Annals of the American Thoracic Society*, 14(6):844–850, 2017. doi: doi.org/10.1513/AnnalsATS.201609-728PS.
- [20] C. Summers, N. R. Singh, J. F. White, I. M. Mackenzie, A. Johnston, C. Solanki, K. K. Balan, A. M. Peters, and E. R. Chilvers. Pulmonary retention of primed neutrophils: a novel protective host response, which is impaired in the acute respiratory distress syndrome. *Thorax*, 69(7):623–629, Jul 2014. doi: doi.org/10.1136/thoraxjnl-2013-204742.
- [21] Fraser R. Millar, Charlotte Summers, Mark J. Griffiths, Mark R. Toshner, and Alastair G. Proudfoot. The pulmonary endothelium in acute respiratory distress syn-

- drome: Insights and therapeutic opportunities. *Thorax*, 71(5):462–473, 2016. doi: doi.org/10.1136/thoraxjnl-2015-207461.
- [22] R. P. Marshall, G. Bellin, S. Webb, A. Puddicombe, N. Goldsack, R. J. McAnulty, and G. J. Laurent. Fibroproliferation occurs early in the acute respiratory distress syndrome and impacts on outcome. *Am. J. Respir. Crit. Care Med.*, 162(5):1783–1788, 2000. doi: 10.1164/ajrccm.162.5.2001061.
- [23] E. Wesley Ely, Ayumi Shintani, Brenda Truman, Theodore Speroff, Sharon M. Gordon, Frank E. Harrell, Jr, Sharon K. Inouye, Gordon R. Bernard, and Robert S. Dittus. Delirium as a Predictor of Mortality in Mechanically Ventilated Patients in the Intensive Care Unit. *JAMA*, 291(14):1753–1762, 04 2004. ISSN 0098-7484. doi: doi.org/10.1001/jama.291.14.1753. URL <https://doi.org/10.1001/jama.291.14.1753>.
- [24] Laurent Papazian, Jean-Marie Forel, Arnaud Gacouin, Christine Penot-Ragon, Gilles Perrin, Anderson Loundou, Samir Jaber, Jean-michel Arnal, Didier Perez, Jean-Marie Seghboyen, Jean-Michel Constantin, Pierre Courant, Jean-Yves Lefrant, Claude Guérin, Gwenaél Prat, Sophie Morange, Antoine Roch, and ACURASYS Study Investigators. Neuromuscular blockers in early acute respiratory distress syndrome. *The New England Journal of Medicine*, 363(12):1107–16, 2010. doi: doi.org/10.1056/NEJMoa1005372.
- [25] Claude Guérin, Jean Reignier, Jean-Christophe Richard, Pascal Beuret, Arnaud Gacouin, Thierry Boulain, Emmanuelle Mercier, Michel Badet, Alain Mercat, Olivier Baudin, Marc Clavel, Delphine Chatellier, Samir Jaber, Sylvène Rosselli, Jordi Mancebo, Michel Sirodot, Gilles Hilbert, Christian Bengler, Jack Richet, Marc Gainnier, Frédérique Bayle, Gael Bourdin, Véronique Leray, Raphael Girard, Loredana Baboi, and Louis Ayzac. Prone Positioning in Severe Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 368(23):2159–2168, 2013. doi: doi.org/10.1056/NEJMoa1214103.
- [26] Jesús Villar, Carlos Ferrando, Domingo Martínez, Alfonso Ambrós, Tomás Muñoz, Juan A Soler, Gerardo Aguilar, Francisco Alba, Elena González-Higueras, Luís A Conesa, Carmen Martín-Rodríguez, Francisco J Díaz-Domínguez, Pablo Serna-Grande, Rosana Rivas, José Ferreres, Javier Belda, Lucía Capilla, Alec Tallet, José M Añón, Rosa L Fernández, Jesús M González-Martín, Gerardo Aguilar, Francisco Alba, Julián Álvarez, Alfonso Ambrós, José M. Añón, María J. Asensio, Javier Belda, Jesús Blanco, Marisa Blasco, Lucía Cachafeiro, Rafael del Campo, Lucía

- Capilla, José A. Carbonell, Nieves Carbonell, Agustín Cariñena, Demetrio Carriedo, Mario Chico, Luís A. Conesa, Ruth Corpas, Javier Cuervo, Francisco J. Díaz-Domínguez, Cristina Domínguez-Antelo, Lorena Fernández, Rosa L. Fernández, Carlos Ferrando, José Ferreres, Eneritz Gamboa, Elena González-Higueras, Raúl I. González-Luengo, Jesús M. González-Martín, Domingo Martínez, Carmen Martín-Rodríguez, Tomás Muñoz, Ramón Ortiz Díaz-Miguel, Raquel Pérez-González, Ana M. Prieto, Isidro Prieto, Rosana Rivas, Leticia Rojas-Viguera, Miguel A. Romera, Jesús Sánchez-Ballesteros, José M. Segura, Pablo Serna-Grande, Ainhoa Serrano, Rosario Solano, Juan A. Soler, Marina Soro, Alec Tallet, and Jesús Villar. Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. *The Lancet Respiratory Medicine*, 8(3):267 – 276, 2020. doi: doi.org/10.1016/S2213-2600(19)30417-5.
- [27] The ARDS Network Authors for the ARDS Network. Ketoconazole for Early Treatment of Acute Lung Injury and Acute Respiratory Distress Syndrome. *JAMA*, 283(15):1995, 2003. doi: doi.org/10.1001/jama.283.15.1995.
- [28] The ARDS Network Authors for the ARDS Network. Placebo-controlled Trial of Lisofylline for Early Treatment of Acute Lung Injury and Acute Respiratory Distress Syndrome Network Participants. *Critical Care Medicine*, 30(1):1–6, 2002. doi: doi.org/10.1097/00003246-200201000-00001.
- [29] The ARDS Network Authors for the ARDS Network. Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 351(4):327–336, 2004. doi: doi.org/10.1056/NEJMoa032193.
- [30] Robert W Taylor, Janice L Zimmerman, Phillip R Straube, Gerard J Criner, Kathleen M Kelly, Thomas C Smith, and Robert J Small. Low-Dose Inhaled Nitric Oxide in Patients with Acute Lung Injury. *JAMA*, 291:1603–1609, 2004. doi: doi.org/10.1097/01.sa.0000144222.87401.0e.
- [31] The ARDS Network Authors for the ARDS Network. Efficacy and Safety of Corticosteroids for Persistent Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 354(16):1671–1684, 2006. doi: doi.org/10.1056/NEJMoa051693.
- [32] Herbert P Wiedemann, Arthur P Wheeler, Gordon R Bernard, B Taylor Thompson, Douglas Hayden, Ben DeBoisblanc, Alfred F Connors, R Duncan Hite, and Andrea L Harabin. Comparison of two fluid-management strategies in acute lung injury. *The*

- New England Journal of Medicine*, 354(24):2564–75, 2006. doi: doi.org/10.1056/NEJMoa062200.
- [33] Todd W. Rice, Arthur P. Wheeler, B. Taylor Thompson, Bennett P. DeBoisblanc, Jay Steingrub, and Peter Rock. Enteral omega-3 fatty acid, γ -linolenic acid, and antioxidant supplementation in acute lung injury. *JAMA*, 306(14):1574–1581, 2011. doi: doi.org/10.1001/jama.2011.1435.
- [34] Michael A. Matthay, Roy G. Brower, Shannon Carson, Ivor S. Douglas, Mark Eisner, Duncan Hite, Steven Holets, Richard H. Kallet, Kathleen D. Liu, Neil MacIntyre, Marc Moss, David Schoenfeld, Jay Steingrub, and B. Taylor Thompson. Randomized, placebo-controlled clinical trial of an aerosolized β 2-agonist for treatment of acute lung injury. *American Journal of Respiratory and Critical Care Medicine*, 184(5): 561–568, 2011. doi: doi.org/10.1164/rccm.201012-2090OC.
- [35] Fang Gao Smith, Gavin D. Perkins, Simon Gates, Duncan Young, Daniel F. McAuley, William Tunnicliffe, Zahid Khan, and Sarah E. Lamb. Effect of intravenous β -2 agonist treatment on clinical outcomes in acute respiratory distress syndrome (BALTI-2): A multicentre, randomised controlled trial. *The Lancet*, 379(9812):229–235, 2012. doi: doi.org/10.1016/S0140-6736(11)61623-1.
- [36] Todd W. Rice, Arthur P. Wheeler, B. Taylor Thompson, Jay Steingrub, R. Duncan Hite, Marc Moss, Alan Morris, Ning Dong, and Peter Rock. Initial trophic vs full enteral feeding in patients with acute lung injury: The EDEN randomized trial. *JAMA*, 307(8):795–803, 2012. doi: doi.org/10.1001/jama.2012.137.
- [37] Danny F. McAuley, John G. Laffey, Cecilia M. O’Kane, Gavin D. Perkins, Brian Mullan, T. John Trinder, Paul Johnston, Philip A. Hopkins, Andrew J. Johnston, Cliona McDowell, and Christine McNally. Simvastatin in the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 371(1):1695–703, 2014. doi: doi.org/10.1097/00000542-199405000-00004.
- [38] Duncan Young, Sarah E. Lamb, Sanjoy Shah, Iain MacKenzie, William Tunnicliffe, R Lall, Kathy Rowan, and Brian H Cuthbertson. High-Frequency Oscillation for Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 368(9): 806–13, 2013. doi: doi.org/10.1056/NEJMoa1215716.
- [39] Niall D. Ferguson, Deborah Cook, Gordon Guyatt, Sangeeta Mehta, Lori Hand, Peggy Austin, Qi Zhou, Andrea Matte, Stephen D Walter, Francois Lamontagne, John T

- Granton, Yaseen M. Arabi, Alejandro C Arroliga, Thomas E. Stewart, Arthur S. Slutsky, and Maureen O. Meade. High frequency oscillation in early acute respiratory distress syndrome. *The New England Journal of Medicine*, 368(3):795–805, 2013. doi: doi.org/10.1056/NEJMoA1215554.
- [40] The ARDS Network Authors for the ARDS Network. Rosuvastatin for Sepsis-Associated Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, 370(23):2191–2200, 2014. doi: doi.org/10.1056/nejmoa1401520.
- [41] Daryl J. Kor, Rickey E. Carter, Pauline K. Park, Emir Festic, Valerie M. Banner-Goodspeed, Richard Hinds, Daniel Talmor, Ognjen Gajic, Lorraine B. Ware, and Michelle Ng Gong. Effect of aspirin on development of ARDS in at-risk patients presenting to the emergency department the LIPS-a randomized clinical trial. *JAMA*, 315(22):2406–2414, 2016. doi: doi.org/10.1001/jama.2016.6330.
- [42] Daniel F. McAuley, LJ Mark Cross, Umar Hamid, Evie Gardner, J. Stuart Elborn, Kathy M. Cullen, Ahilanandan Dushianthan, Michael PW Grocott, Michael A. Matthay, and Cecilia M. O’Kane. Keratinocyte growth factor for the treatment of the acute respiratory distress syndrome (KARE): a randomised, double-blind, placebo-controlled phase 2 trial. *The Lancet Respiratory Medicine*, 5(6):484–491, 2017. doi: doi.org/10.1016/S2213-2600(17)30171-6.
- [43] Alain Combes, David Hajage, G Capellier, A Demoule, S Lavoue, C. Guervilly, D. Da Silva, L. Zafrani, P. Tirot, B Veber, E. Maury, B. Levy, Y. Cohen, C. Richard, P. Kalfon, L. Bouadma, H. Mehdaoui, G. Beduneau, G. Lebreton, L. Brochard, N.D. Ferguson, E Fan, . A.S Slutsky, D. Brodie, and Mercat A. Extracorporeal membrane oxygenation for acute respiratory distress syndrome in adults. *The New England Journal of Medicine*, 378(21):1965–75, 2018. doi: doi.org/10.1097/NCI.0b013e31828a09ff.
- [44] PETAL Clinical Trials Network (NHLBI). Early Neuromuscular Blockade in the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, pages 1–12, 2019. doi: doi.org/10.1056/nejmoa1901686.
- [45] P. Asfar, F. Meziani, J. F. Hamel, F. Grelon, B. Megarbane, N. Anguel, J. P. Mira, P. F. Dequin, S. Gergaud, N. Weiss, F. Legay, and Y. Le Tulzo. High versus low blood-pressure target in patients with septic shock. *The New England Journal of Medicine*, 370(17):1583–1593, Apr 2014. doi: doi.org/10.1056/NEJMoA1312173.
- [46] A. C. Gordon, G. D. Perkins, M. Singer, D. F. McAuley, R. M. Orme, S. Santhakumaran, A. J. Mason, M. Cross, F. Al-Beidh, J. Best-Lane, D. Brealey, C. L. Nutt, J. J.

- McNamee, H. Reschreiter, A. Breen, K. D. Liu, and D. Ashby. Levosimendan for the Prevention of Acute Organ Dysfunction in Sepsis. *The New England Journal of Medicine*, 375(17):1638–1648, 10 2016. doi: doi.org/10.1056/NEJMoa1609409.
- [47] A. Zarbock, J. A. Kellum, C. Schmidt, H. Van Aken, C. Wempe, H. Pavenstadt, A. Boanta, J. Gerss, and M. Meersch. Effect of Early vs Delayed Initiation of Renal Replacement Therapy on Mortality in Critically Ill Patients With Acute Kidney Injury: The ELAIN Randomized Clinical Trial. *JAMA*, 315(20):2190–2199, 2016. doi: doi.org/10.1001/jama.2016.5828.
- [48] H. Thiele, U. Zeymer, F. J. Neumann, M. Ferenc, H. G. Olbrich, J. Hausleiter, G. Richardt, M. Hennersdorf, K. Empen, G. Fuernau, S. Desch, I. Eitel, R. Hambrecht, J. Fuhrmann, M. Bohm, H. Ebelt, S. Schneider, G. Schuler, and K. Werdan. Intraaortic balloon support for myocardial infarction with cardiogenic shock. *The New England Journal of Medicine*, 367(14):1287–1296, Oct 2012. doi: doi.org/10.1056/NEJMoa1208410.
- [49] V. Marco Ranieri, B. Taylor Thompson, Philip S Barie, JF Dhainaut, Ivor S Douglas, Simon Finfer, Bengt Gårdlund, John C Marshall, Andrew Rhodes, Mark D Williams, and Prowess-shock Study Group. Drotrecogin Alfa (Activated) in Adults with Septic Shock. *The New England Journal of Medicine*, 366(22):2055–2064, 2012. doi: doi.org/10.1056/NEJMoa1202290.
- [50] Anders Perner, Nicolai Haase, Anne Guttormse, Jyrki Tenhunen, Kristian R Madsen, Morten H Møller, Jeanie M Elkjær, Jonas Nielsen, Lasse H Andersen, Lars B Holst, Per Winkel, and Jørn Wetterslev. Hydroxyethyl Starch 130/0.42 versus Ringer’s Acetate in Severe Sepsis. *The New England Journal of Medicine*, 367(2):124–34, 2012. doi: doi.org/10.1056/NEJMoa1204242.
- [51] Raiko Blondonnet, Jean-michel Constantin, Vincent Sapin, and Matthieu Jabaudon. A Pathophysiologic Approach to Biomarkers in Acute Respiratory Distress Syndrome. *Disease Markers*, 2016:1–20, 2016. doi: doi.org/10.1155/2016/3501373.
- [52] Racheal G. Akwii, Md S. Sajib, Fatema T. Zahra, and Constantinos M. Mikelis. Role of angiopoietin-2 in vascular physiology and pathophysiology. *Cells*, 8(5), 2019. ISSN 2073-4409. doi: doi.org/10.3390/cells8050471. URL <https://www.mdpi.com/2073-4409/8/5/471>.
- [53] L. B. Ware, Z. Zhao, T. Koyama, R. M. Brown, M. W. Semler, D. R. Janz, A. K. May, R. D. Fremont, M. A. Matthay, M. J. Cohen, and C. S. Calfee. Derivation

- and validation of a two-biomarker panel for diagnosis of ARDS in patients with severe traumatic injuries. *Trauma Surg Acute Care Open*, 2(1):e000121, 2017. doi: doi.org/10.1136/tsaco-2017-000121.
- [54] A. Agrawal, M. A. Matthay, K. N. Kangelaris, J. Stein, J. C. Chu, B. M. Imp, A. Cortez, J. Abbott, K. D. Liu, and C. S. Calfee. Plasma angiopoietin-2 predicts the onset of acute lung injury in critically ill patients. *American Journal of Respiratory and Critical Care Medicine*, 187(7):736–742, Apr 2013. doi: doi.org/10.1164/rccm.201208-1460OC.
- [55] Lorraine B Ware, Mark D Eisner, B Taylor Thompson, Polly E Parsons, and Michael A Matthay. Significance of von willebrand factor in septic and nonseptic patients with acute lung injury. *American journal of respiratory and critical care medicine*, 170(7):766—772, October 2004. ISSN 1073-449X. doi: doi.org/10.1164/rccm.200310-1434oc. URL <https://doi.org/10.1164/rccm.200310-1434OC>.
- [56] Lorraine B Ware, Michael A Matthay, Polly E Parsons, B Taylor Thompson, James L Januzzi, Mark D Eisner, and National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome Clinical Trials Network. Pathogenetic and prognostic significance of altered coagulation and fibrinolysis in acute lung injury/acute respiratory distress syndrome. *Critical care medicine*, 35(8):1821—1828, August 2007. ISSN 0090-3493. doi: doi.org/10.1097/01.ccm.0000221922.08878.49. URL <https://europepmc.org/articles/PMC2764536>.
- [57] M. L. Terpstra, J. Aman, G. P. van Nieuw Amerongen, and A. B. Groeneveld. Plasma biomarkers for acute respiratory distress syndrome: a systematic review and meta-analysis. *Critical Care Medicine*, 42(3):691–700, Mar 2014. doi: doi.org/10.1097/01.ccm.0000435669.60811.24.
- [58] Lorraine B. Ware, Tatsuki Koyama, D. Dean Billheimer, William Wu, Gordon R. Bernard, B. Taylor Thompson, Roy G. Brower, Theodore J. Standiford, Thomas R. Martin, Michael A. Matthay, and G. R. Bernard. Prognostic and pathogenetic value of combining clinical and biochemical indices in patients with acute lung injury. *Chest*, 137(2):288–296, 2010. doi: doi.org/10.1378/chest.09-1484.
- [59] Zhiguo Zhao, Nancy Wickersham, Kirsten N Kangelaris, Addison K May, Gordon R Bernard, Michael A Matthay, Carolyn S Calfee, Tatsuki Koyama, and Lorraine B Ware. External validation of a biomarker and clinical prediction model for hospital mortality in acute respiratory distress syndrome. *Intensive Care Medicine*, 43(8): 1123–1131, 2017. doi: doi.org/10.1007/s00134-017-4854-5.

- [60] Lorraine B. Ware, Tatsuki Koyama, Zhiguo Zhao, David R. Janz, Nancy Wickersham, Gordon R. Bernard, Addison K. May, Carolyn S. Calfee, and Michael A. Matthay. Biomarkers of lung epithelial injury and inflammation distinguish severe sepsis patients with acute respiratory distress syndrome. *Critical Care*, 17(5):1, 2013. doi: doi.org/10.1186/cc13080.
- [61] Carolyn S. Calfee, David R. Janz, Gordon R. Bernard, Addison K. May, Kirsten N. Kangelaris, Michael A. Matthay, and Lorraine B. Ware. Distinct molecular phenotypes of direct vs indirect ARDS in single-center and multicenter studies. *Chest*, 147(6): 1539–1548, 2015. doi: doi.org/10.1378/chest.14-2454.
- [62] Timothy E. Sweeney, Neal J. Thomas, Judie A. Howrylak, Hector R. Wong, Angela J. Rogers, and Purvesh Khatri. Multicohort analysis of whole-blood gene expression data does not form a robust diagnostic for acute respiratory distress syndrome. *Critical Care Medicine*, 46(2):244–251, 2018. doi: doi.org/10.1097/CCM.0000000000002839.
- [63] Kirsten Neudoerffer Kangelaris, Arun Prakash, Kathleen D. Liu, Bradley Aouizerat, Prescott G. Woodruff, David J. Erle, Angela Rogers, Eric J. Seeley, Jeffrey Chu, Tom Liu, Thomas Osterberg-Deiss, Hanjing Zhuo, Michael A. Matthay, and Carolyn S. Calfee. Increased expression of neutrophil-related genes in patients with early sepsis-induced ARDS. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 308(11):L1102–L1113, 2015. doi: doi.org/10.1152/ajplung.00380.2014.
- [64] JA Howrylak, T Dolinay, L Lucht, Z Wang, David Christiani, Jigme Sethi, Eric Xing, Michael Donahoe, and Augustine Choi. Discovery of the gene signature for acute lung injury in patients with sepsis. *Physiological Genomics*, 37:133–139, 2009. doi: doi.org/10.1152/physiolgenomics.90275.2008.
- [65] Y. Chen, J. X. Shi, X. F. Pan, J. Feng, and H. Zhao. DNA microarray-based screening of differentially expressed genes related to acute lung injury and functional analysis. *European Review for Medical and Pharmacological Sciences*, 17(8):1044–1050, 2013.
- [66] Tamás Dolinay, Young Sam Kim, Judie Howrylak, Gary M Hunninghake, Chang Hyeok An, Laura Fredenburgh, Anthony F Massaro, Angela Rogers, Lee Gazourian, Kiichi Nakahira, Jeffrey A Haspel, Roberto Landazury, Sabitha Epanapally, Jason D Christie, Nuala J Meyer, Lorraine B Ware, David C Christiani, Stefan W Ryter, Rebecca M Baron, and Augustine M.K. Choi. Inflammasome-regulated cytokines are critical mediators of acute lung injury. *American Jour-*

- nal of Respiratory and Critical Care Medicine*, 185(11):1225–1234, 2012. doi: doi.org/10.1164/rccm.201201-0003OC.
- [67] Jatinder K. Juss, David House, Augustin Amour, Malcolm Begg, Jurgen Herre, Daniel M.L. Storisteanu, Kim Hoenderdos, Glyn Bradley, Mark Lennon, Charlotte Summers, Edith M. Hessel, Alison Condliffe, and Edwin R. Chilvers. Acute respiratory distress syndrome neutrophils have a distinct phenotype and are resistant to phosphoinositide 3-kinase inhibition. *American Journal of Respiratory and Critical Care Medicine*, 194(8):961–973, 2016. doi: doi.org/10.1164/rccm.201509-1818OC.
- [68] Jason D. Christie, Mark M. Wurfel, Rui Feng, Grant E. O’Keefe, Jonathan Bradfield, Lorraine B. Ware, David C. Christiani, Carolyn S. Calfee, Mitchell J. Cohen, Michael Matthay, Nuala J. Meyer, Cecilia Kim, Mingyao Li, Joshua Akey, Kathleen C. Barnes, Jonathan Sevransky, Paul N. Lanken, Addison K. May, Richard Aplenc, James P. Maloney, and Hakon Hakonarson. Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. *PLoS ONE*, 7(1): 1–10, 2012. doi: doi.org/10.1371/journal.pone.0028268.
- [69] Christian Bime, Nima Pouladi, Saad Sammani, Ken Batai, Nancy Casanova, Tong Zhou, Carrie L. Kempf, Xiaoguang Sun, Sara M. Camp, Ting Wang, Rick A. Kittles, Yves A. Lussier, Tiffanie K. Jones, John P. Reilly, Nuala J. Meyer, Jason D. Christie, Jason H. Karnes, Manuel Gonzalez-Garay, David C. Christiani, Charles R. Yates, Mark M. Wurfel, Gianfranco U. Meduri, and Joe G.N. Garcia. Genome-wide association study in African Americans with acute respiratory distress syndrome identifies the selectin P ligand gene as a risk factor. *American Journal of Respiratory and Critical Care Medicine*, 197(11):1421–1432, 2018. doi: doi.org/10.1164/rccm.201705-0961OC.
- [70] Natalia Hernandez-Pacheco, Beatriz Guillen-Guio, Marialbert Acosta-Herrera, Maria Pino-Yanes, Almudena Corrales, Alfonso Ambrós, Leonor Nogales, Arturo Muriel, Elena González-Higueras, Francisco J Diaz-Dominguez, Elizabeth Zavala, Javier Belda, Shwu-Fan Ma, Jesús Villar, Carlos Flores, Jesús Villar, Rosa L Fernández, Carlos Flores, Maria Pino-Yanes, Marialbert Acosta-Herrera, Lina Pérez-Méndez, Almudena Corrales, Elena Espinosa, David Domínguez, Jesús Blanco, Arturo Muriel, Victor Sagredo, Juan C Ballesteros, Alfonso Ambrós, Rafael Cdel Campo, Francisco Gandía, Leonor Nogales, Rafael Fernández, Carles Subirá, Aurora Baluja, José M Añón, Elena González, Rosario Solano, Demetrio Carriedo, Francisco J Diaz-Dominguez, Ramón Adalia, Elizabeth Zavala, and the GEN-SEP Network. A vascular

- endothelial growth factor receptor gene variant is associated with susceptibility to acute respiratory distress syndrome. *Intensive Care Med. Exp.*, 6(1):16, 2018. doi: doi.org/10.1186/s40635-018-0181-6.
- [71] G D Perkins, J Roberts, D F McAuley, L Armstrong, A Millar, F Gao, and D R Thickett. Regulation of vascular endothelial growth factor bioactivity in patients with acute lung injury. *Thorax*, 60(2):153–158, 2005. doi: doi.org/10.1136/thx.2004.027912.
- [72] John P Reilly, Fan Wang, Tiffanie K Jones, Jessica A Palakshappa, Brian J Anderson, Michael G S Shashaty, Thomas G Dunn, Erik D Johansson, Thomas R Riley, Brian Lim, Jason Abbott, Caroline A G Ittner, Edward Cantu, Xihong Lin, Carmen Mikacenic, Mark M Wurfel, David C Christiani, Carolyn S Calfee, Michael A Matthay, Jason D Christie, Rui Feng, and Nuala J Meyer. Plasma angiopoietin-2 as a potential causal marker in sepsis-associated ARDS development: evidence from Mendelian randomization and mediation analysis. *Intensive Care Medicine*, 44(11):1849–1858, 2018. doi: doi.org/10.1007/s00134-018-5328-0.
- [73] Michelle N Gong, Wei Zhou, Paige L Williams, Taylor B Thompson, Lucille Pothier, and David C Christiani. Polymorphisms in the mannose binding lectin-2 gene and acute respiratory distress syndrome*. *Crit. Care Med.*, 35(1), 2007. doi: doi.org/10.1097/01.CCM.0000251132.10689.F3.
- [74] John P A Ioannidis, Evangelia E Ntzani, Thomas A Trikalinos, and Despina G Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nat. Genet.*, 29(3):306–309, 2001. ISSN 1546-1718. doi: doi.org/10.1038/ng749. URL <https://doi.org/10.1038/ng749>.
- [75] Patrick F Sullivan. Spurious Genetic Associations. *Biol. Psychiatry*, 61(10):1121–1126, may 2007. ISSN 0006-3223. doi: doi.org/10.1016/j.biopsych.2006.11.010. URL <https://doi.org/10.1016/j.biopsych.2006.11.010>.
- [76] Colin R. Cooke, Chirag V. Shah, Robert Gallop, Scarlett Bellamy, Marek Ancukiewicz, Mark D. Eisner, Paul N. Lanken, A. Russell Localio, and Jason D. Christie. A simple clinical predictive index for objective estimates of mortality in acute lung injury. *Critical Care Medicine*, 37(6):1913–1920, 2009. doi: doi.org/10.1097/CCM.0b013e3181a009b4.
- [77] Lisa M. Brown, Carolyn S. Calfee, Michael A. Matthay, Roy G. Brower, B. Taylor Thompson, and William Checkley. A simple classification model for hospital

- mortality in patients with acute lung injury managed with lung protective ventilation. *Critical Care Medicine*, 39(12):2645–2651, 2011. doi: doi.org/10.1097/CCM.0b013e3182266779.
- [78] J. Villar, R. L. Fernandez, A. Ambros, L. Parra, J. Blanco, A. M. Dominguez-Berrot, J. M. Gutierrez, L. Blanch, J. M. Anon, C. Martin, F. Prieto, J. Collado, L. Perez-Mendez, and R. M. Kacmarek. A clinical classification of the acute respiratory distress syndrome for predicting outcome and guiding medical therapy. *Critical Care Medicine*, 43(2):346–353, Feb 2015. doi: doi.org/10.1097/CCM.0000000000000703.
- [79] Lieuwe D. Bos, Olaf L. Cremer, David S.Y. Ong, Eliana B. Caser, Carmen S.V. Barbas, Jesus Villar, Robert M. Kacmarek, and Marcus J. Schultz. External validation confirms the legitimacy of a new clinical classification of ARDS for predicting outcome. *Intensive Care Medicine*, 41(11):2004–2005, 2015. doi: doi.org/10.1007/s00134-015-3992-x.
- [80] Chen Yu Wang, Carolyn S. Calfee, Devon W. Paul, David R. Janz, Addison K. May, Hanjing Zhuo, Gordon R. Bernard, Michael A. Matthay, Lorraine B. Ware, and Kirsten Neudoerffer Kangelaris. One-year mortality and predictors of death among hospital survivors of acute respiratory distress syndrome. *Intensive Care Medicine*, 40(3):388–396, 2014. doi: doi.org/10.1007/s00134-013-3186-3.
- [81] L. D. J. Bos, B. P. Scicluna, D. S. Y. Ong, O. Cremer, T. van der Poll, and M. J. Schultz. Understanding Heterogeneity in Biologic Phenotypes of Acute Respiratory Distress Syndrome by Leukocyte Expression Profiles. *Am. J. Respir. Crit. Care Med.*, 200(1): 42–50, 2019. doi: doi.org/10.1164/rccm.201809-1808OC.
- [82] L. D. Bos, L. R. Schouten, L. A. van Vught, M. A. Wiewel, D. S. Y. Ong, O. Cremer, A. Artigas, I. Martin-Loeches, A. J. Hoogendijk, T. van der Poll, J. Horn, N. Juffermans, C. S. Calfee, M. J. Schultz, J. F. Frencken, M. Bonten, P. M. C. Klein Klouwenberg, R. T. M. van Hooijdonk, M. A. Huson, M. Straat, E. Witteveen, G. J. Glas, L. Wieske, B. P. Scicluna, and H. Belkasim-Bohoudi. Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis. *Thorax*, 72(10):876–883, 2017. doi: doi.org/10.1136/thoraxjnl-2016-209719.
- [83] Matthew J Loza, Ratko Djukanovic, Kian Fan Chung, Daniel Horowitz, Keying Ma, Patrick Branigan, Elliot S Barnathan, Vedrana S Susulic, Philip E Silkoff, Peter J Sterk, and Frédéric Baribaud. Validated and longitudinally stable asthma phenotypes

- based on cluster analysis of the ADEPT study. *Respiratory Research*, 17(1):1–21, 2016. doi: doi.org/10.1186/s12931-016-0482-9.
- [84] V Siroux, X Basagan, A Boudier, I Pin, J. Garcia-Aymerich, A Vesin, R Slama, D Jarvis, J M Anto, F Kauffmann, and J Sunyer. Identifying adult asthma phenotypes using a clustering approach. *European Respiratory Journal*, 38(2):310–317, 2011. doi: doi.org/10.1183/09031936.00120810.
- [85] P. Haldar, I. D. Pavord, D. E. Shaw, M. A. Berry, M. Thomas, C. E. Brightling, A. J. Wardlaw, and R. H. Green. Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine*, 178(3):218–224, Aug 2008. doi: doi.org/10.1164/rccm.200711-1754OC.
- [86] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D’Agostino, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, B. Gaston, N. N. Jarjour, R. Sorkness, W. J. Calhoun, K. F. Chung, S. A. Comhair, R. A. Dweik, E. Israel, S. P. Peters, W. W. Busse, S. C. Erzurum, and E. R. Bleeker. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *American Journal of Respiratory and Critical Care Medicine*, 181(4):315–323, Feb 2010. doi: doi.org/10.1164/rccm.200906-0896OC.
- [87] Sally E Wenzel. Asthma Phenotypes: The Evolution from Clinical to Molecular Approaches. *Nature Medicine*, 18(5):716–725, 2012. doi: doi.org/10.1038/nm.2678.
- [88] Rebecca Howard, Magnus Rattray, Mattia Prosperi, and Adnan Custovic. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Current Allergy and Asthma Reports*, 15(7), 2015. doi: doi.org/10.1007/s11882-015-0542-0.
- [89] T. S. Hinks, T. Brown, L. C. Lau, H. Rupani, C. Barber, S. Elliott, J. A. Ward, J. Ono, S. Ohta, K. Izuhara, R. Djukanovi?, R. J. Kurukulaarachy, A. Chauhan, and P. H. Howarth. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3-like protein 1. *J. Allergy Clin. Immunol.*, 138(1):61–75, 2016. doi: doi.org/10.1016/j.jaci.2015.11.020.
- [90] Shaheenah Dawood, Rong Hu, Michelle D Homes, Laura C. Collins, Stuart J. Schnitt, James Connolly, Graham A Colditz, and Rulla M Tamimi. Defining breast cancer prognosis based on molecular phenotypes: Results from a large cohort study. *Breast Cancer Research and Treatment*, 126(1):185–192, 2011. doi: doi.org/10.1007/s10549-010-1113-7.

- [91] Rulla M Tamimi, Heather J Baer, Jonathan Marotti, Mark Galan, Laurie Galaburda, Yineng Fu, Anne C Deitz, James L Connolly, Stuart J Schnitt, Graham A Colditz, and Laura C Collins. Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Research*, 10(4):1–9, 2008. doi: doi.org/10.1186/bcr2128.
- [92] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and L. J. Esserman. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharmacol. Ther.*, 86(1):97–100, 2009. doi: doi.org/10.1038/clpt.2009.68.
- [93] S. Das and A. W. Lo. Re-inventing drug development: A case study of the I-SPY 2 breast cancer clinical trials program. *Contemp Clin Trials*, 62:168–174, 2017. doi: doi.org/10.1016/j.cct.2017.09.002.
- [94] N. Gobat, J. Amuasi, Y. Yazdanpanah, L. Sigfid, H. Davies, J. P. Byrne, G. Carson, C. Butler, A. Nichol, and H. Goossens. Advancing preparedness for clinical research during infectious disease epidemics. *ERJ Open Res*, 5(2), 2019. doi: doi.org/10.1183/23120541.00227-201.
- [95] Annemarie B Docherty, Ewen M Harrison, Christopher A Green, Hayley E Hardwick, Riinu Pius, Lisa Norman, Karl A Holden, Jonathan M Read, Frank Dondelinger, Gail Carson, Laura Merson, James Lee, Daniel Plotkin, Louise Sigfrid, Sophie Halpin, Clare Jackson, Carrol Gamble, Peter W Horby, Jonathan S Nguyen-Van-Tam, Antonia Ho, Clark D Russell, Jake Dunning, Peter JM Openshaw, J Kenneth Baillie, and Malcolm G Semple. Features of 20,133 UK patients in hospital with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ*, 369, 2020. doi: doi.org/10.1136/bmj.m1985.
- [96] B. P. Scicluna, L. A. van Vught, A. H. Zwinderman, M. A. Wiewel, E. E. Davenport, K. L. Burnham, P. Nurnberg, M. J. Schultz, J. Horn, O. L. Cremer, M. J. Bonten, C. J. Hinds, H. R. Wong, J. C. Knight, T. van der Poll, F. M. de Beer, L. D. J. Bos, J. F. Frencken, M. E. Koster-Brouwer, K. van de Groep, D. M. Verboom, G. J. Glas, R. T. M. van Hooijdonk, A. J. Hoogendijk, M. A. Huson, P. M. Klein Klouwenberg, D. S. Y. Ong, L. R. A. Schouten, M. Straat, E. Witteveen, and L. Wieske. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *The Lancet Respiratory Medicine*, 5(10):816–826, 10 2017. doi: doi.org/10.1016/S2213-2600(17)30294-1.

- [97] Emma E. Davenport, Katie L. Burnham, Jayachandran Radhakrishnan, Peter Humburg, Paula Hutton, Tara C. Mills, Anna Rautanen, Anthony C. Gordon, Christopher Garrard, Adrian V.S. Hill, Charles J. Hinds, and Julian C. Knight. Genomic landscape of the individual host response and outcomes in sepsis: A prospective cohort study. *The Lancet Respiratory Medicine*, 4(4):259–271, 2016. doi: doi.org/10.1016/S2213-2600(16)00046-1.
- [98] Katie L. Burnham, Emma E. Davenport, Jayachandran Radhakrishnan, Peter Humburg, Anthony C. Gordon, Paula Hutton, Eduardo Svoren-Jabalera, Christopher Garrard, Adrian V.S. Hill, Charles J. Hinds, and Julian C. Knight. Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 196(3):328–339, 2017. doi: doi.org/10.1164/rccm.201608-1685OC.
- [99] Anna Rautanen, Tara C. Mills, Anthony C. Gordon, Paula Hutton, Michael Steffens, Rosamond Nuamah, Jean Daniel Chiche, Tom Parks, Stephen J. Chapman, Emma E. Davenport, Katherine S. Elliott, Julian Bion, Peter Lichtner, Thomas Meitinger, Thomas F. Wienker, Mark J. Caulfield, Charles Mein, Frank Bloos, Ilona Bobek, Paolo Cotogni, Vladimir Sramek, Silver Sarapuu, Makbule Kobilay, V. Marco Ranieri, Jordi Rello, Gonzalo Sirgo, Yoram G. Weiss, Stefan Russwurm, E. Marion Schneider, Konrad Reinhart, Paul A.H. Holloway, Julian C. Knight, Chris S. Garrard, James A. Russell, Keith R. Walley, Frank Stüber, Adrian V.S. Hill, and Charles J. Hinds. Genome-wide association study of survival from sepsis due to pneumonia: An observational cohort study. *The Lancet Respiratory Medicine*, 3(1):53–60, 2015. doi: doi.org/10.1016/S2213-2600(14)70290-5.
- [100] Anthony C. Gordon, Alexina J. Mason, Neeraja Thirunavukkarasu, Gavin D. Perkins, Maurizio Cecconi, Magda Cepkova, David G. Pogson, Hollmann D. Aya, Aisha Anjum, Gregory J. Frazier, Shalini Santhakumaran, Deborah Ashby, Stephen J. Brett, and for the VANISH Investigators. Effect of Early Vasopressin vs Norepinephrine on Kidney Failure in Patients With Septic Shock: The VANISH Randomized Clinical Trial. *JAMA*, 316(5):509–518, 08 2016. doi: doi.org/10.1001/jama.2016.10485.
- [101] David B. Antcliffe, Katie L. Burnham, Farah Al-Beidh, Shalini Santhakumaran, Stephen J. Brett, Charles J. Hinds, Deborah Ashby, Julian C. Knight, and Anthony C. Gordon. Transcriptomic signatures in sepsis and a differential response to steroids. from the vanish randomized trial. *American Journal of Respiratory and Critical Care Medicine*, 199(8):980–986, 2019. doi: doi.org/10.1164/rccm.201807-1419OC.

- [102] C. W. Seymour, J. N. Kennedy, S. Wang, C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, D. T. Huang, J. A. Kellum, Q. Mi, S. M. Opal, V. Talisa, T. van der Poll, S. Visweswaran, Y. Vodovotz, J. C. Weiss, D. M. Yealy, S. Yende, and D. C. Angus. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA*, 2019. doi: doi.org/10.1001/jama.2019.5791.
- [103] Mihael Ankerst, Markus M. Breunig, Hans peter Kriegel, and Jörg Sander. OPTICS: Ordering Points to Identify the Clustering Structure. In *ACM SIGMOD Record*, volume 28, pages 49–60, New York, NY, USA, 1999. Association for Computing Machinery. doi: doi.org/10.1145/304181.304187.
- [104] Carolyn S. Calfee, Kevin Delucchi, Polly E. Parsons, B. Taylor Thompson, Lorraine B. Ware, and Michael A. Matthay. Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *The Lancet Respiratory Medicine*, 2(8):611–620, 2014. doi: doi.org/10.1016/S2213-2600(14)70097-9.
- [105] Katie R. Famous, Kevin Delucchi, Lorraine B. Ware, Kirsten N. Kangelaris, Kathleen D. Liu, B. Taylor Thompson, and Carolyn S. Calfee. Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *American Journal of Respiratory and Critical Care Medicine*, 195(3):331–338, 2017. doi: doi.org/10.1164/rccm.201603-0645OC.
- [106] Carolyn S. Calfee, Kevin L. Delucchi, Pratik Sinha, Michael A. Matthay, Jonathan Hackett, Manu Shankar-Hari, Cliona McDowell, John G. Laffey, Cecilia M. O’Kane, and Daniel F. McAuley. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. *The Lancet Respiratory Medicine*, 6(9):691–698, 2018. doi: doi.org/10.1016/S2213-2600(18)30177-2.
- [107] Kevin Delucchi, Katie R Famous, Lorraine B Ware, Polly E Parsons, B Taylor Thompson, and Carolyn S Calfee. Stability of ards subphenotypes over time in two randomised controlled trials. *Thorax*, 73(5):439–445, 2018. doi: doi.org/10.1136/thoraxjnl-2017-211090.
- [108] Pratik Sinha, Kevin L. Delucchi, B. Taylor Thompson, Daniel F. McAuley, Michael A. Matthay, and Carolyn S. Calfee. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS)

- study. *Intensive Care Medicine*, 44(11):1859–1869, 2018. doi: doi.org/10.1007/s00134-018-5378-3.
- [109] John P Reilly, Scarlett Bellamy, Michael G.S. Shashaty, Robert Gallop, Nuala J Meyer, Paul N Lanken, Sandra Kaplan, Daniel N Holena, Addison K May, Lorraine B Ware, and Jason D Christie. Heterogeneous phenotypes of acute respiratory distress syndrome after major trauma. *Annals of the American Thoracic Society*, 11(5):728–736, 2014. doi: doi.org/10.1513/AnnalsATS.201308-280OC.
- [110] Jake Dunning, Simon Blankley, Long T. Hoang, Mike Cox, Christine M. Graham, Philip L. James, Chloe I. Bloom, Damien Chaussabel, Jacques Banchereau, Stephen Brett, Miriam F. Moffatt, Anne O’Garra, Peter J.M. Openshaw, and MOSAIC Investigators. Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. *Nature Immunology*, 19(6): 625–635, 2018. doi: doi.org/10.1038/s41590-018-0111-5.
- [111] Murali Shyamsundar, Scott T. W. McKeown, Cecilia M. O’Kane, Thelma R. Craig, Vanessa Brown, David R. Thickett, Michael A. Matthay, Clifford C. Taggart, Janne T. Backman, J. Stuart Elborn, and Daniel F. McAuley. Simvastatin decreases lipopolysaccharide-induced pulmonary inflammation in healthy volunteers. *American Journal of Respiratory and Critical Care Medicine*, 179(12):1107–1114, 2009. doi: doi.org/10.1164/rccm.200810-1584OC.
- [112] Kenneth Rockwood, Xiaowei Song, Chris MacKnight, Howard Bergman, David B. Hogan, Ian McDowell, and Arnold Mitnitski. A global clinical measure of fitness and frailty in elderly people. *CMAJ*, 173(5):489–495, 2005. doi: doi.org/10.1503/cmaj.050051.
- [113] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014. URL <http://www.jstatsoft.org/v61/i06/>.
- [114] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: doi.org/10.18637/jss.v045.i03.
- [115] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*, 6(1):21689, 2016. doi: doi.org/10.1038/srep21689.

- [116] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: doi.org/10.1007/BF01908075.
- [117] Du, P., Kibbe, W.A., Lin, and S.M. nuid: A universal naming schema of oligonucleotides for illumina, affymetrix, and other microarrays. *Biology Direct*, 2007.
- [118] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi: doi.org/10.1093/nar/gkv007.
- [119] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22):14031–14036, 2002. doi: doi.org/10.1073/pnas.222164199.
- [120] Ramona Schmid, Patrick Baum, Carina Ittrich, Katrin Fundel-Clemens, Wolfgang Huber, Benedikt Brors, Roland Eils, Andreas Weith, Detlev Mennerich, and Karsten Quast. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics*, 11(1), 2010. doi: doi.org/10.1186/1471-2164-11-349.
- [121] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, mar 2012. doi: doi.org/10.1093/bioinformatics/bts034.
- [122] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 04 2006. doi: doi.org/10.1093/biostatistics/kxj037.
- [123] Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLOS ONE*, 6(2):1–10, 02 2011. doi: doi.org/10.1371/journal.pone.0017238.
- [124] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4: Article17, 2005. doi: doi.org/10.2202/1544-6115.1128.
- [125] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8(1):22, 2007. doi: doi.org/10.1186/1471-2105-8-22.

- [126] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, Dec 2008. doi: 10.1186/1471-2105-9-559.
- [127] Christopher J. Walsh, Jane Batt, Margaret S. Herridge, Sunita Mathur, Gary D. Bader, Pingzhao Hu, and Claudia C. Dos Santos. Transcriptomic analysis reveals abnormal muscle repair and remodeling in survivors of critical illness with sustained weakness. *Nature Scientific Reports*, 6:1–9, 2016. doi: doi.org/10.1038/srep29334.
- [128] Jerome Friedman, Trevor Hastie, and Robert Tibishirani. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, NY, second edition, 2008. ISBN 978-0-387-84858-7. doi: doi.org/10.1007/978-0-387-84858-7.
- [129] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL <http://www.stats.ox.ac.uk/pub/MASS4>.
- [130] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2019. URL <https://CRAN.R-project.org/package=caret>. R package version 6.0-84.
- [131] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, 10(1):1523, 2019. doi: doi.org/10.1038/s41467-019-09234-6.
- [132] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma’ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1): 128, 2013. doi: doi.org/10.1186/1471-2105-14-128.
- [133] Benoît De Hertogh, Bertrand De Meulder, Fabrice Berger, Michael Pierre, Eric Bareke, Anthoula Gaigneaux, and Eric Depiereux. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC Bioinformatics*, 11:17, jan 2010. doi: doi.org/10.1186/1471-2105-11-17.
- [134] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL <https://www.R-project.org/>.

- [135] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. URL <https://CRAN.R-project.org/package=dplyr>. R package version 1.0.0.
- [136] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL <https://ggplot2.tidyverse.org>.
- [137] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.
- [138] Alboukadel Kassambara and Marcin Kosinski. *survminer: Drawing Survival Curves using 'ggplot2'*, 2019. URL <https://CRAN.R-project.org/package=survminer>. R package version 0.4.4.
- [139] Masaaki Horikoshi and Yuan Tang. *ggfortify: Data Visualization Tools for Statistical Analysis Results*, 2018. URL <https://CRAN.R-project.org/package=ggfortify>.
- [140] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2019. URL <https://CRAN.R-project.org/package=ggpubr>. R package version 0.2.1.
- [141] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2019. URL <https://CRAN.R-project.org/package=cowplot>. R package version 1.0.0.
- [142] J.J. Allaire, Christopher Gandrud, Kenton Russell, and CJ Yetman. *networkD3: D3 JavaScript Network Graphs from R*, 2017. URL <https://CRAN.R-project.org/package=networkD3>. R package version 0.4.
- [143] RStudio and Inc. *htmltools: Tools for HTML*, 2017. URL <https://CRAN.R-project.org/package=htmltools>. R package version 0.3.6.
- [144] Katherine C. Verbist, David L. Rose, Charles J. Cole, Mary B. Field, and Kimberly D. Klonowski. Il-15 participates in the respiratory innate immune response to influenza virus infection. *PLOS ONE*, 7:1–11, 05 2012. doi: doi.org/10.1371/journal.pone.0037539.
- [145] Liel Cohen, Andrew Fiore-Gartland, Adrienne G Randolph, Angela Panoskaltsis-Mortari, Sook-San Wong, Jacqui Ralston, Timothy Wood, Ruth Seeds, Q Sue Huang, Richard J Webby, Paul G Thomas, and Tomer Hertz. A Modular Cytokine Analysis Method Reveals Novel Associations With Clinical Phenotypes and Identifies Sets of Co-signaling Cytokines Across Influenza Natural Infection Cohorts and Healthy

- Controls. *Frontiers in immunology*, 10:1338, jun 2019. doi: doi.org/10.3389/fimmu.2019.01338.
- [146] Fengyi Wan and Michael J Lenardo. The nuclear signaling of NF-kappaB: current knowledge, new insights, and future perspectives. *Cell Res.*, 20(1):24–33, jan 2010. doi: doi.org/10.1038/cr.2009.137.
- [147] Ricardo Grieshaber-Bouyer and Peter A. Nigrovic. Neutrophil heterogeneity as therapeutic opportunity in immune-mediated disease. *Frontiers in Immunology*, 10: 346, 2019. doi: doi.org/10.3389/fimmu.2019.00346.
- [148] Ivan Nagaev, Maria Bokarewa, Andrej Tarkowski, and Ulf Smith. Human resistin is a systemic immune-derived proinflammatory cytokine targeting both leukocytes and adipocytes. *PLOS ONE*, 1(1):1–9, 12 2006. doi: doi.org/10.1371/journal.pone.0000031.
- [149] Jillian R Richter, Jeffrey M Sutton, Ritha M Belizaire, Lou Ann Friend, Rebecca M Schuster, Taylor A Johannigman, Steven G Miller, Alex B Lentsch, and Timothy A Pritts. Macrophage-derived chemokine (CCL22) is a novel mediator of lung inflammation following hemorrhage and resuscitation. *Shock*, 42(6):525–531, dec 2014. doi: doi.org/10.1097/SHK.0000000000000253.
- [150] Rupak Mukhopadhyay, Jie Jia, Abul Arif, Partho Sarothi Ray, and Paul L Fox. The GAIT system: a gatekeeper of inflammatory gene expression. *Trends in Biochemical Sciences*, 34(7):324–331, 2009. doi: https://doi.org/10.1016/j.tibs.2009.03.004.
- [151] Y Arimori, R Nakamura, H Yamada, K Shibata, N Maeda, T Kase, and Y Yoshikai. Type I Interferon Plays Opposing Roles in Cytotoxicity and Interferon- γ Production by Natural Killer and CD8⁺ T Cells after Influenza A Virus Infection in Mice. *Journal of Innate Immunity*, 6(4):456–466, 2014. doi: doi.org/10.1159/000356824.
- [152] Abul Arif, Jie Jia, Robyn A. Moodt, Paul E. DiCorleto, and Paul L. Fox. Phosphorylation of glutamyl-prolyl trna synthetase by cyclin-dependent kinase 5 dictates transcript-selective translational control. *Proceedings of the National Academy of Sciences*, 108(4):1415–1420, 2011. doi: doi.org/10.1073/pnas.1011275108.
- [153] Eun-Young Lee, Hyun-Cheol Lee, Hyun-Kwan Kim, Song Yee Jang, Seong-Jun Park, Yong-Hoon Kim, Jong Hwan Kim, Jungwon Hwang, Jae-Hoon Kim, Tae-Hwan Kim, Abul Arif, Seon-Young Kim, Young-Ki Choi, Cheolju Lee, Chul-Ho Lee, Jae U Jung,

- Paul L Fox, Sunghoon Kim, Jong-Soo Lee, and Myung Hee Kim. Infection-specific phosphorylation of glutamyl-prolyl tRNA synthetase induces antiviral immunity. *Nature Immunology*, 17(11):1252–1262, 2016. doi: doi.org/10.1038/ni.3542.
- [154] Kohsaku Uetani, Miki Hiroi, Tadamichi Meguro, Hiroshi Ogawa, Toshinori Kamisako, Yoshihiro Ohmori, and Serpil C. Erzurum. Influenza a virus abrogates ifn- γ response in respiratory epithelial cells by disruption of the jak/stat pathway. *European Journal of Immunology*, 38(6):1559–1573, 2008. doi: https://doi.org/10.1002/eji.200737045.
- [155] Anzheng Nie, Bao Sun, Zhihui Fu, and Dongsheng Yu. Roles of aminoacyl-tRNA synthetases in immune regulation and immune diseases. *Cell Death & Disease*, 10(12):901, 2019. doi: doi.org/10.1038/s41419-019-2145-5.
- [156] Peter Walter and David Ron. The unfolded protein response: From stress pathway to homeostatic regulation. *Science*, 334(6059):1081–1086, 2011. doi: doi.org/10.1126/science.1209038. URL https://science.sciencemag.org/content/334/6059/1081.
- [157] Violeta Chitu and E Richard Stanley. Colony-stimulating factor-1 in immunity and inflammation. *Current Opinion in Immunology*, 18(1):39–48, 2006. doi: https://doi.org/10.1016/j.coi.2005.11.006.
- [158] Huiting Su, Ning Na, Xiaodong Zhang, and Yong Zhao. The biological function and significance of CD74 in immune diseases. *Inflammation Research*, 66(3):209–216, 2017. doi: doi.org/10.1007/s00011-016-0995-1.
- [159] Lan Ma and Gang Pei. β -arrestin signaling and regulation of transcription. *Journal of Cell Science*, 120(2):213–218, 01 2007. doi: doi.org/10.1242/jcs.03338.
- [160] Peter ten Dijke, Marie-José Goumans, and Evangelia Pardali. Endoglin in angiogenesis and vascular diseases. *Angiogenesis*, 11(1):79–89, 2008. doi: doi.org/10.1007/s10456-008-9101-9.
- [161] Vanessa Boury Faiotto, Daniel Franci, Rodolfo Monteiro Enz Hubert, Gleice Regina de Souza, Maiara Marx Luz Fiusa, Bidossessi Wilfried Hounkpe, Thiago Martins Santos, Marco Antonio Carvalho-Filho, and Erich Vinicius De Paula. Circulating levels of the angiogenesis mediators endoglin, hb-egf, bmp-9 and fgf-2 in patients with severe sepsis and septic shock. *Journal of Critical Care*, 42:162–167, 2017. doi: https://doi.org/10.1016/j.jcrc.2017.07.034.

- [162] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: doi.org/10.1073/pnas.122653799.
- [163] Manuel Ramos-Casals, Pilar Brito-Zerón, Armando López-Guillermo, Munther A Khamashta, and Xavier Bosch. Adult haemophagocytic syndrome. *Lancet*, 383(9927): 1503–1516, apr 2014. doi: [doi.org/10.1016/S0140-6736\(13\)61048-X](https://doi.org/10.1016/S0140-6736(13)61048-X).
- [164] Kris Bauchmuller, Jessica J Manson, Rachel Tattersall, Michael Brown, Christopher McNamara, Mervyn Singer, and Stephen J Brett. Haemophagocytic lymphohistiocytosis in adult critical care. *Journal of the Intensive Care Society*, 21(3):256–268, 2020. doi: doi.org/10.1177/1751143719893865.
- [165] Jan-Inge Henter, AnnaCarin Horne, Maurizio Aricó, R. Maarten Egeler, Alexandra H. Filipovich, Shinsaku Imashuku, Stephan Ladisch, Ken McClain, David Webb, Jacek Winiarski, and Gritta Janka. Hlh-2004: Diagnostic and therapeutic guidelines for hemophagocytic lymphohistiocytosis. *Pediatric Blood & Cancer*, 48(2):124–131, 2007. doi: <https://doi.org/10.1002/pbc.21039>.
- [166] Gabriela M. Wochnik, Joëlle Rüegg, G. Alexander Abel, Ulrike Schmidt, Florian Holsboer, and Theo Rein. Fk506-binding proteins 51 and 52 differentially regulate dynein interaction and nuclear translocation of the glucocorticoid receptor in mammalian cells. *Journal of Biological Chemistry*, 280(6):4609–4616, 2005. doi: doi.org/10.1074/jbc.M407498200.
- [167] Susana Orozco and Andrew Oberst. RIPK3 in cell death and inflammation: the good, the bad, and the ugly. *Immunol. Rev.*, 277(1):102–112, may 2017. doi: doi.org/10.1111/imr.12536.
- [168] Erika L. Pearce and Edward J. Pearce. Metabolic pathways in immune cell activation and quiescence. *Immunity*, 38(4):633 – 643, 2013. ISSN 1074-7613. doi: <https://doi.org/10.1016/j.immuni.2013.04.005>. URL <http://www.sciencedirect.com/science/article/pii/S1074761313001581>.
- [169] Nyall R. London, Weiquan Zhu, Fernando A. Bozza, Matthew C.P. Smith, Daniel M. Greif, Lise K. Sorensen, Luming Chen, Yuuki Kaminoh, Aubrey C. Chan, Samuel F. Passi, Craig W. Day, Dale L. Barnard, Guy A. Zimmerman, Mark A. Krasnow, and Dean Y. Li. Targeting Robo4-dependent slit signaling to survive the cytokine storm in sepsis and influenza. *Science Translational Medicine*, 2(23), 2010. doi: doi.org/10.1126/scitranslmed.3000678.

- [170] Elena E. Gorbunova, Irina N. Gavrilovskaya, and Erich R. Mackow. Slit2-Robo4 receptor responses inhibit ANDV directed permeability of human lung microvascular endothelial cells. *Antiviral Res.*, 99(2):108–112, 2013. doi: doi.org/10.1016/j.antiviral.2013.05.004.
- [171] Jie Weng, Xiaoming Zhou, Hui Xie, Ye Gao, Zhiyi Wang, and Yuqiang Gong. Slit2/Robo4 signaling pathway modulates endothelial hyper-permeability in a two-event in vitro model of transfusion-related acute lung injury. *Blood Cells, Molecules, and Diseases*, 76(November):7–12, 2019. doi: doi.org/10.1016/j.bcmed.2018.11.003.
- [172] Bryan Ericksen, Zhibin Wu, Wuyuan Lu, and Robert I Lehrer. Antibacterial activity and specificity of the six human alpha-defensins. *Antimicrob. Agents Chemother.*, 49(1):269–275, jan 2005. doi: doi.org/10.1128/AAC.49.1.269-275.2005.
- [173] Alexander R Moschen, Timon E Adolph, Romana R Gerner, Verena Wieser, and Herbert Tilg. Lipocalin-2: A master mediator of intestinal and metabolic inflammation. *Trends Endocrinol. Metab.*, 28(5):388–397, may 2017. doi: doi.org/10.1016/j.tem.2017.01.003.
- [174] K R Wasiluk, K M Skubitz, and B H Gray. Comparison of granule proteins from human polymorphonuclear leukocytes which are bactericidal toward *Pseudomonas aeruginosa*. *Infect. Immun.*, 59(11):4193–4200, nov 1991. doi: doi.org/10.1128/IAI.59.11.4193-4200.1991.
- [175] P W Gray, G Flaggs, S R Leong, R J Gumina, J Weiss, C E Ooi, and P Elsbach. Cloning of the cDNA of a human neutrophil bactericidal protein. structural and functional correlations. *Journal of Biological Chemistry*, 264(16):9505–9, 1989. URL <http://www.jbc.org/content/264/16/9505.abstract>.
- [176] Christopher J Kuckleburg, Sarah B Tilkens, Sentot Santoso, and Peter J Newman. Proteinase 3 contributes to transendothelial migration of NB1-positive neutrophils. *J. Immunol.*, 188(5):2419–2426, mar 2012. doi: doi.org/10.4049/jimmunol.1102540.
- [177] R. Z. Topic and S. Dodig. Eosinophil cationic protein—current concepts and controversies. *Biochem Med (Zagreb)*, 21(2):111–121, 2011. doi: doi.org/10.11613/BM.2011.019.
- [178] J E Gabay, R W Scott, D Campanelli, J Griffith, C Wilde, M N Marra, M Seeger, and C F Nathan. Antibiotic proteins of human polymorphonuclear leukocytes. *Proc. Natl. Acad. Sci. U. S. A.*, 86(14):5610–5614, jul 1989. doi: doi.org/10.1073/pnas.86.14.5610.

- [179] Chao Liu, Zhaojun Xu, Dipika Gupta, and Roman Dziarski. Peptidoglycan recognition proteins: A novel family of four human innate immunity pattern recognition molecules. *Journal of Biological Chemistry*, 276(37):34686–34694, 2001. doi: doi.org/10.1074/jbc.M105566200.
- [180] Abderr azzaq Belaaouaj, Kwang Sik Kim, and Steven D. Shapiro. Degradation of outer membrane Protein A in Escherichia coli killing by neutrophil elastase. *Science*, 289(5482):1185–1187, 2000. doi: doi.org/10.1126/science.289.5482.1185.
- [181] Wayne Bellamy, Mitsunori Takase, Koji Yamauchi, Hiroyuki Wakabayashi, Kouzou Kawase, and Mamoru Tomita. Identification of the bactericidal domain of lactoferrin. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1121(1):130 – 136, 1992. doi: https://doi.org/10.1016/0167-4838(92)90346-F.
- [182] Benjamin M Tang, Maryam Shojaei, Sally Teoh, Adrienne Meyers, John Ho, T Blake Ball, Yoav Keynan, Amarnath Pisipati, Aseem Kumar, Damon P Eisen, Kevin Lai, Mark Gillett, Rahul Santram, Robert Geffers, Jens Schreiber, Khyobeni Mozhui, Stephen Huang, Grant P Parnell, Marek Nalos, Monika Holubova, Tracy Chew, David Booth, Anand Kumar, Anthony McLean, and Klaus Schughart. Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection. *Nature Communications*, 10(1):3422, 2019. doi: doi.org/10.1038/s41467-019-11249-y.
- [183] Cyndi Goh, Katie L Burnham, M Azim Ansari, Mariateresa de Cesare, Tanya Golubchik, Paula Hutton, Lauren E Overend, Emma E Davenport, Charles J Hinds, Rory Bowden, and Julian C Knight. Epstein-Barr virus reactivation in sepsis due to community-acquired pneumonia is associated with increased morbidity and an immunosuppressed host transcriptomic endotype. *Scientific Reports*, 10(1):9838, 2020. doi: doi.org/10.1038/s41598-020-66713-3.
- [184] Agnieszka Kaczmarek, Peter Vandenabeele, and Dmitri V. Krysko. Necroptosis: The Release of Damage-Associated Molecular Patterns and Its Physiological Relevance. *Immunity*, 38(2):209–223, feb 2013. ISSN 1074-7613. doi: doi.org/10.1016/j.immuni.2013.02.003.
- [185] J. P. Nicholson, M. R. Wolmarans, and G. R. Park. The role of albumin in critical illness. *Br J Anaesth*, 85(4):599–610, Oct 2000. doi: doi.org/10.1093/bja/85.4.599.

- [186] A. M. Smith and J. A. McCullers. Secondary bacterial infections in influenza virus infection pathogenesis. *Curr Top Microbiol Immunol*, 385:327–356, 2014. doi: doi.org/10.1007/82_2014_394.
- [187] D. A. Papanicolaou, R. L. Wilder, S. C. Manolagas, and G. P. Chrousos. The pathophysiologic roles of interleukin-6 in human disease. *Ann Intern Med*, 128(2):127–137, Jan 1998. doi: doi.org/10.7326/0003-4819-128-2-199801150-00009.
- [188] U. Bali, T. Phillips, H. Hunt, and J. Unitt. FKBP5 mRNA Expression Is a Biomarker for GR Antagonism. *J Clin Endocrinol Metab*, 101(11):4305–4312, 11 2016. doi: doi.org/10.1210/jc.2016-1624.
- [189] N. E. Hammond, A. Corley, and J. F. Fraser. The utility of procalcitonin in diagnosis of H1N1 influenza in intensive care patients. *Anaesth Intensive Care*, 39(2):238–241, Mar 2011. doi: doi.org/10.1177/0310057X1103900213.
- [190] Diego Geroldi, Colomba Falcone, and Enzo Emanuele. Soluble Receptor for Advanced Glycation End Products: From Disease Marker to Potential Therapeutic Target. *Current Medicinal Chemistry*, 13(17):1971–1978, 2006. doi: doi.org/10.2174/092986706777585013.
- [191] Georgios D Kitsios, Libing Yang, Dimitris V Manatakis, Mehdi Nouraie, John Evankovich, William Bain, Daniel G Dunlap, Faraaz Shah, Ian J Barbash, Sarah F Rapport, Yingze Zhang, Rebecca S DeSensi, Nathaniel M Weathington, Bill B Chen, Prabir Ray, Rama K Mallampalli, Panayiotis V Benos, Janet S Lee, Alison Morris, and Bryan J McVerry. Host-response subphenotypes offer prognostic enrichment in patients with or at risk for acute respiratory distress syndrome. *Crit. Care Med.*, 47(12), 2019. doi: doi.org/10.1097/CCM.0000000000004018.
- [192] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1):1017, 2019. doi: doi.org/10.1038/s41467-019-08746-5.
- [193] Emily A Voigt, Diane E Grill, Michael T Zimmermann, Whitney L Simon, Inna G Ovsyannikova, Richard B Kennedy, and Gregory A Poland. Transcriptomic signatures of cellular and humoral immune responses in older adults after seasonal influenza vaccination identified by data-driven clustering. *Scientific Reports*, 8(1):739, 2018. doi: doi.org/10.1038/s41598-017-17735-x.
- [194] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer

- Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 11 2018. doi: doi.org/10.1093/nar/gky1131.
- [195] Jing-Dong J Han, Denis Dupuy, Nicolas Bertin, Michael E Cusick, and Marc Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005. doi: doi.org/10.1038/nbt1116.
- [196] Martin Ester, Hans peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, pages 226–231. AAAI Press, 1996.

Appendices

A Ethical approvals

GAinS

Ethics approval was granted nationally (REC Reference Number 05/MRE00/38 and 08/H0505/78) and for individual participating centres. Written, informed consent was obtained from all patients or a legal representative.

MOSAIC

Registered study number NCT00965354. NHS National Research Ethics Service, Outer West London REC 09/H0709/52, 09/MRE00/67. Additional adult healthy control subjects were recruited as part of a separate study and consented to their samples being used in additional studies (Central London 3 Research Ethics Committee, 09/H0716/41)

HARP-2

Ethics approval was granted nationally (REC name HSC REC B, reference 10/NIR02/36, date of approval 8 Sep 2010).

B Details of microarray experiments

GAINS

Samples were taken for gene expression profiling by rapidly isolating the total blood leucocyte population from whole blood samples (about 10 mL) taken following admission to ICU by use of the LeukoLOCK (Thermo Fisher Scientific, Waltham, MA, USA) depletion filter technology. Total RNA was purified. Illumina Human-HT-12 version 4 Expression BeadChips with 47 231 probes (Illumina, San Diego, CA, USA) was used for genome-wide transcription profiling for the first available sample taken following ICU admission.

MOSAIC

At each time point, 3 mL of whole blood was collected into each of two Tempus tubes (Applied Biosystems/Ambion) by trained research staff following a standard phlebotomy

protocol. Blood was vigorously mixed immediately following collection and was stored at -80°C before RNA extraction. For each patient, the contents of one tube were used for analysis, and the other tube was retained in case of assay failure.

RNA was isolated using 1.5 mL whole blood and the MagMAX-96 Blood RNA Isolation Kit (Applied Biosystems/Ambion), as per the manufacturer's instructions. 250 μg of isolated total RNA was globin-reduced using the GLOBINclear 96-well format kit (Applied Biosystems/Ambion) according to the manufacturer's instructions.

Total and globin-reduced RNA integrity was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). RNA yield was assessed using a NanoDrop8000 spectrophotometer (NanoDrop Products, Thermo Fisher Scientific). High-quality (>6.5 RIN) whole blood RNA was successfully obtained and processed by microarray in all cases. Biotinylated, amplified antisense complementary RNA (cRNA) targets were prepared from 200–250 ng of globin-reduced RNA using the Illumina CustomPrep RNA amplification kit (Applied Biosystems/Ambion).

For each sample, 750 ng of labeled cRNA was hybridized overnight to Illumina Human HT12 V4 BeadChip arrays (Illumina), which contained greater than 47 000 probes. The arrays were washed, blocked, stained and scanned on an Illumina iScan, as per the manufacturer's instructions. GenomeStudio was used to perform quality control and generate signal intensity values.

C Methods for protein biomarker quantifications assays

MOSAIC

Samples were centrifuged at $1000 \times g$ for ten minutes in a cooled centrifuge. A minimum period of thirty minutes from time of phlebotomy had to have passed prior to centrifugation, to ensure clot formation. Several 0.5 – 1.0 mL aliquots of serum supernatant were pipetted into individual, labelled Cryovials and frozen immediately at -80°C . Samples were not allowed to thaw until the day of analysis.

The following panels of mediators were measured using the Mesoscale Discovery (MSD) platform:

- 7-plex: IFN- γ , IL-13, IL-15, IL-17, ITAC, MIG, MIP-1 α
- 10-plex: GM-CSF, IL-1 β , IL-2, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12p70, TNF- α

- 9-plex: Eotaxin-1, Eotaxin-3, MIP-1 β , TARC, IP-10, IL-8, MCP-1, MDC, MCP-4
- 2-plex: IFN- α 2a, IFN- λ (IL-29)
- Single-plex: IFN- β

Inflammatory soluble immune mediator electrochemiluminescence assay analyzed on an MSD SECTOR instrument. For each mediator, a coefficient variation cut-off of 10% was used to set the lower limit of detection. Sample results below the GM-LLOD (geometric mean lower limit of detection) were assigned half the value of the respective GM-LLOD. For each mediator, the calibrator standard curve was plotted by inputting the concentration of each standard in pg/mL into the MSD Discovery Workbench software. The curve was modelled using least squares fitting algorithms, so that signals from samples could be converted into concentrations.

GAinS

In the GAinS study the cytokines were measured using the ProcartaPlex™ Luminex platform (ThermoFischer Scientific, Waltham, MA, USA) following the manufacturer's instructions. In addition to the samples, each 96-well plate contained two blank wells and duplicates of seven gradient dilutions of standards. Samples were randomised between plates to minimize the potential influence of batch effect. Meanwhile serial samples from same patients were kept together.

Only non-haemolytic plasma samples that had never been thawed were used. Samples were prepared according to manufacturer's instructions with no dilution of samples involved. Capture beads were incubated overnight at 4°C. Data was acquired on a Luminex 100 system at the Kennedy Institute of Rheumatology, University of Oxford. Minimum counts triggering a warning message was set to 100 bead reads per bead region.

To correct for background fluorescence, Median Fluorescence Intensity (MFI) was divided by the average MFI of two blank wells on the specific plate. Absolute concentration levels of measured cytokines were determined by mapping MFI back to the plate- and analyte-specific standard curves with R package 'nCal'.

HARP-2

Samples for interleukin 6 and soluble tumour necrosis factor receptor 1 (sTNFr1), Angiopoietin-2, sRAGE, surfactant protein-D and MMP-8 measurement were taken before randomisation.

Plasma from these samples was stored at -80°C. Biomarkers were measured in duplicate with commercially available ELISAs (R&D Systems, Minneapolis, MN, USA).

D Boxplots of protein biomarker concentrations in each cluster

D.1 GAINs

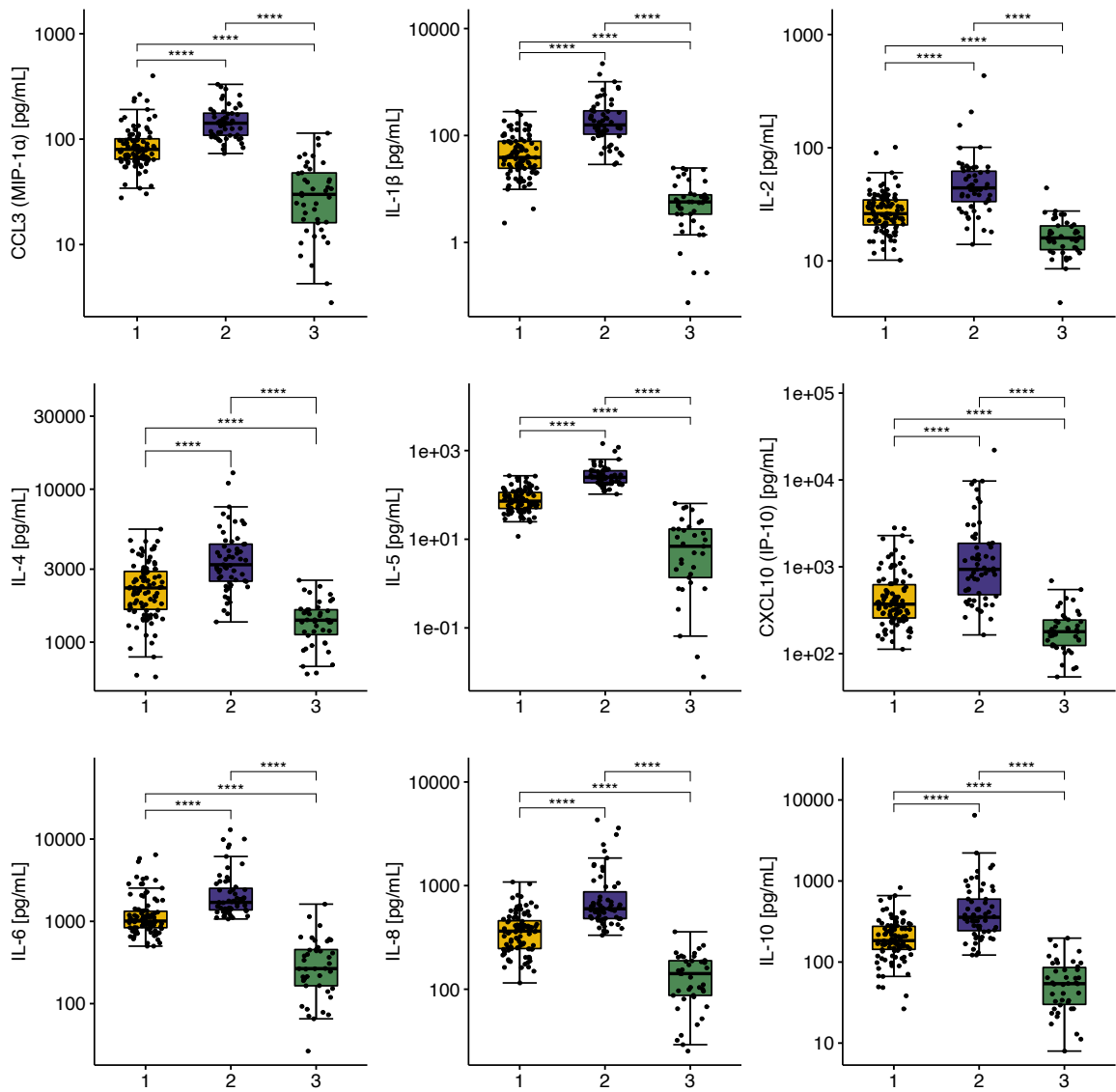


Figure D.1a

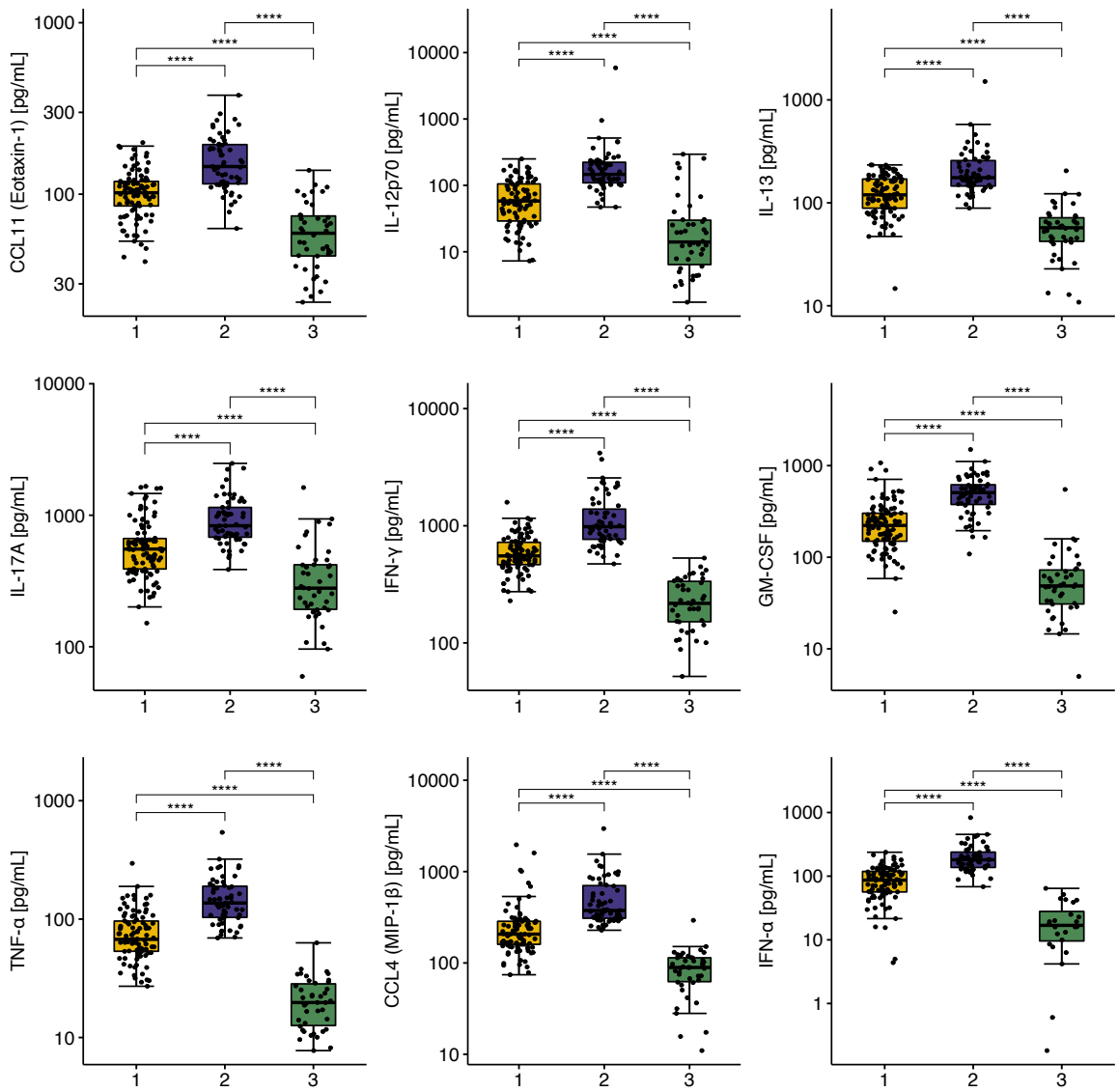
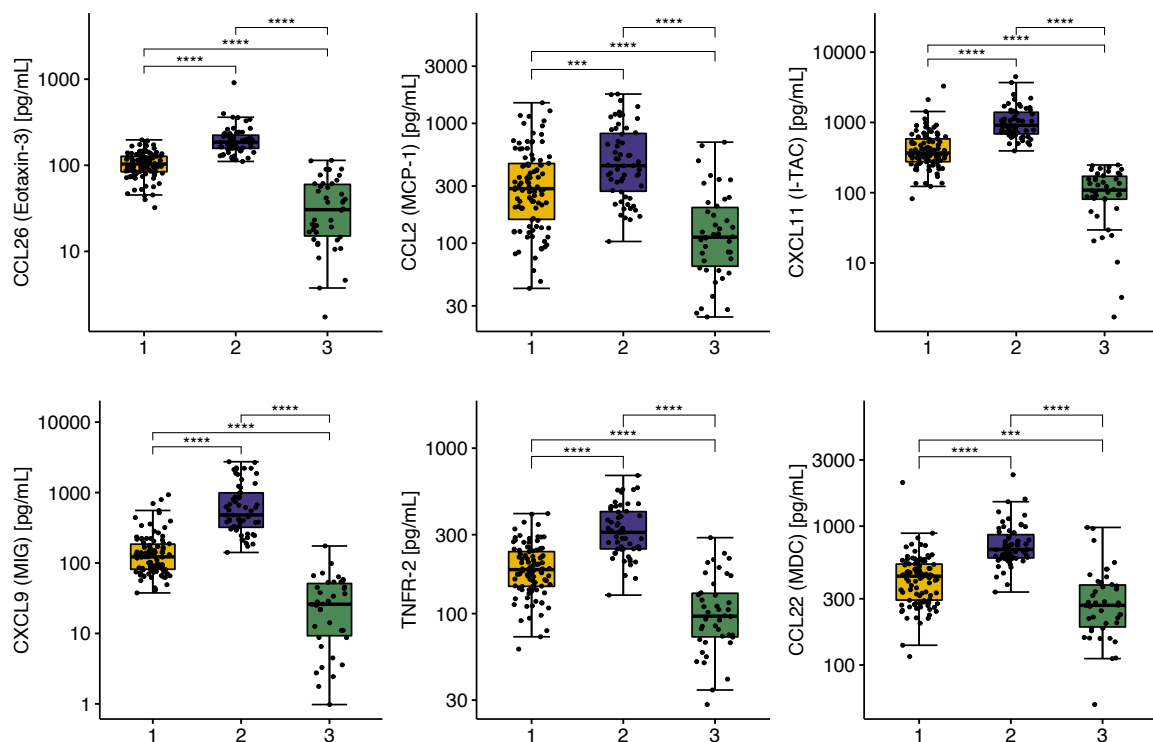


Figure D.1b



(c)

Fig. D.1 Boxplots showing distribution of cytokine and chemokine concentrations in each of the GAIN clusters. The horizontal axes labels and boxplot colours denote the clusters. All comparisons were made using Dunn's test with FDR correction of p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$

D.2 MOSAIC

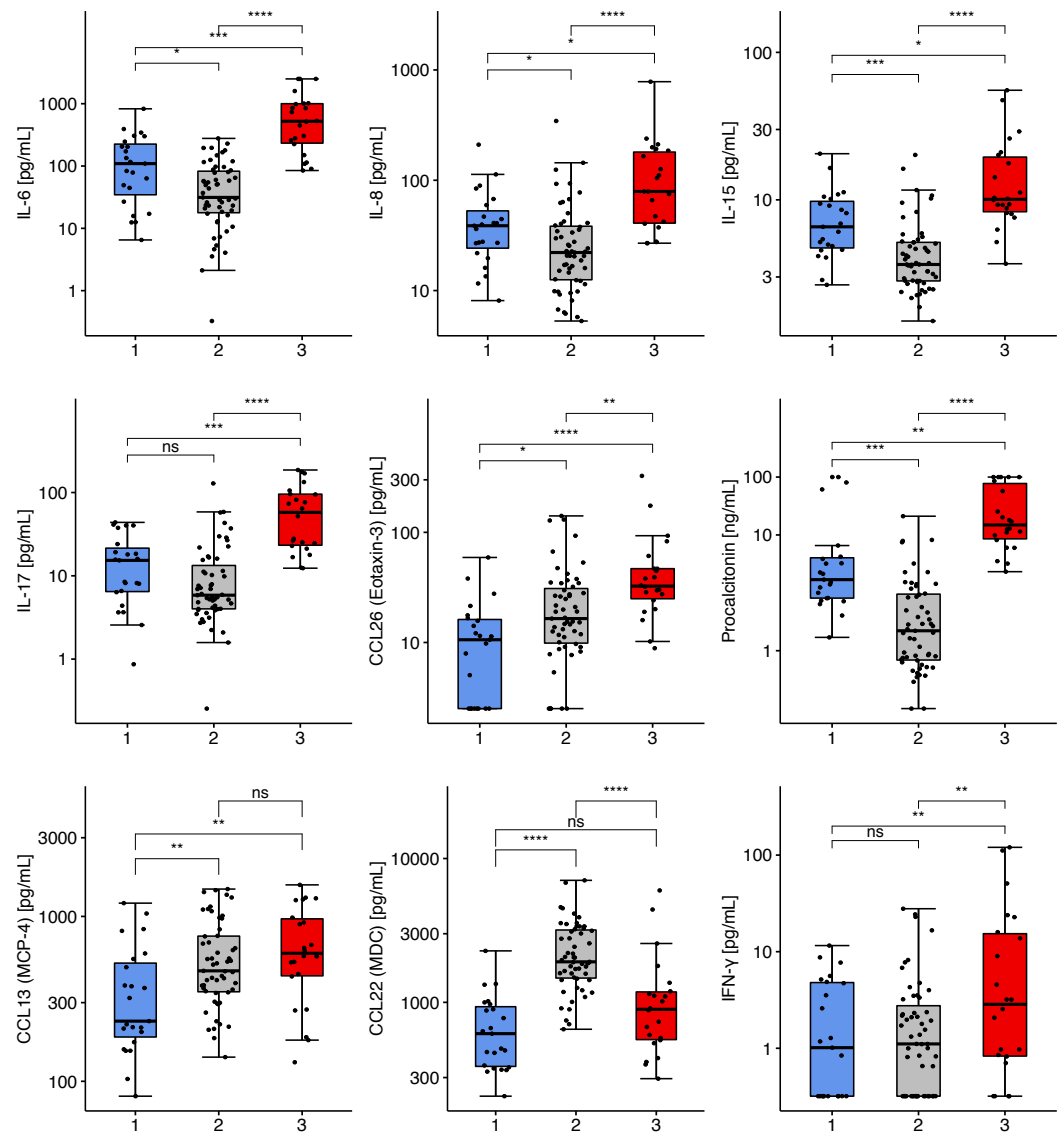


Figure D.2a

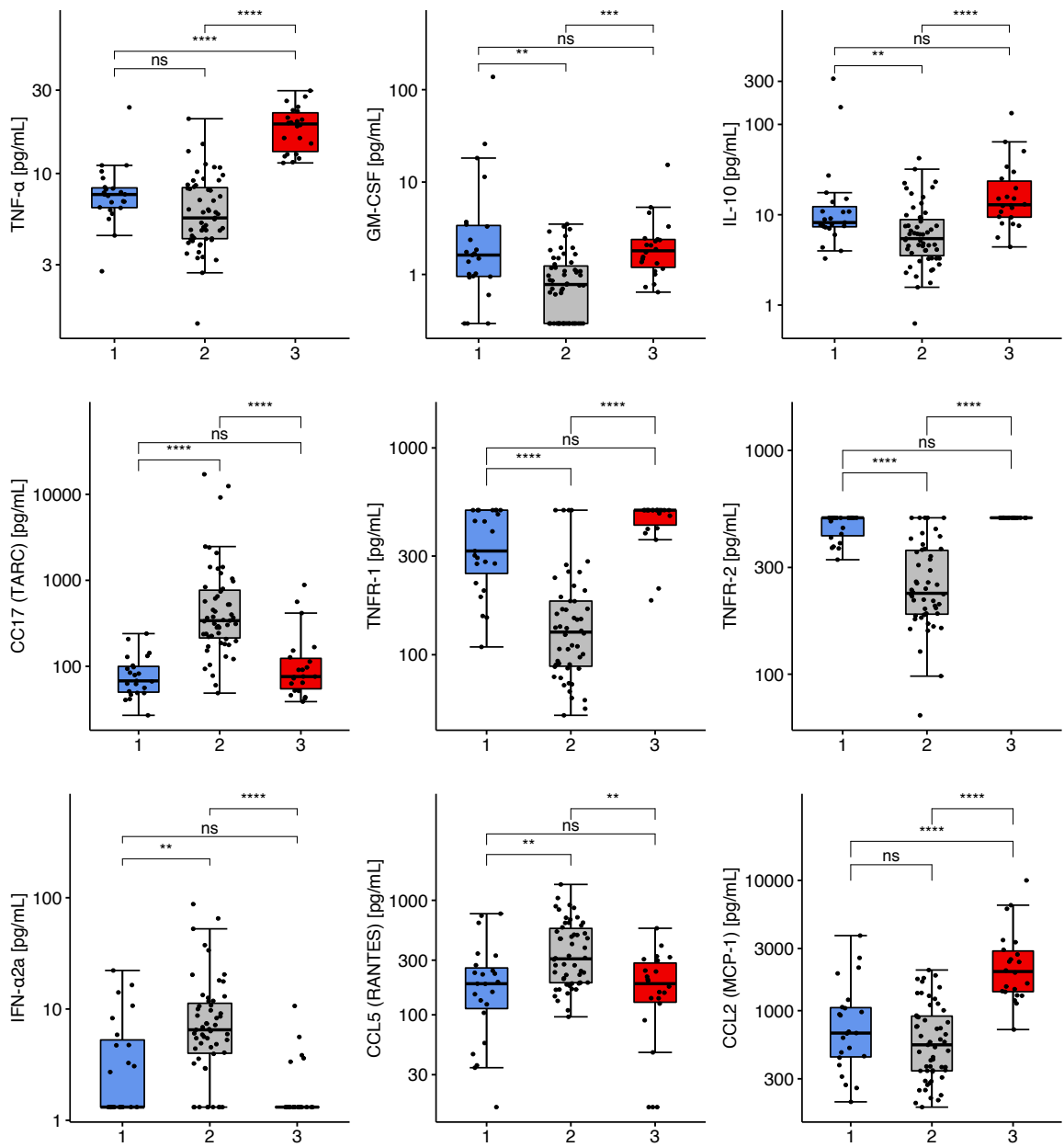


Figure D.2b

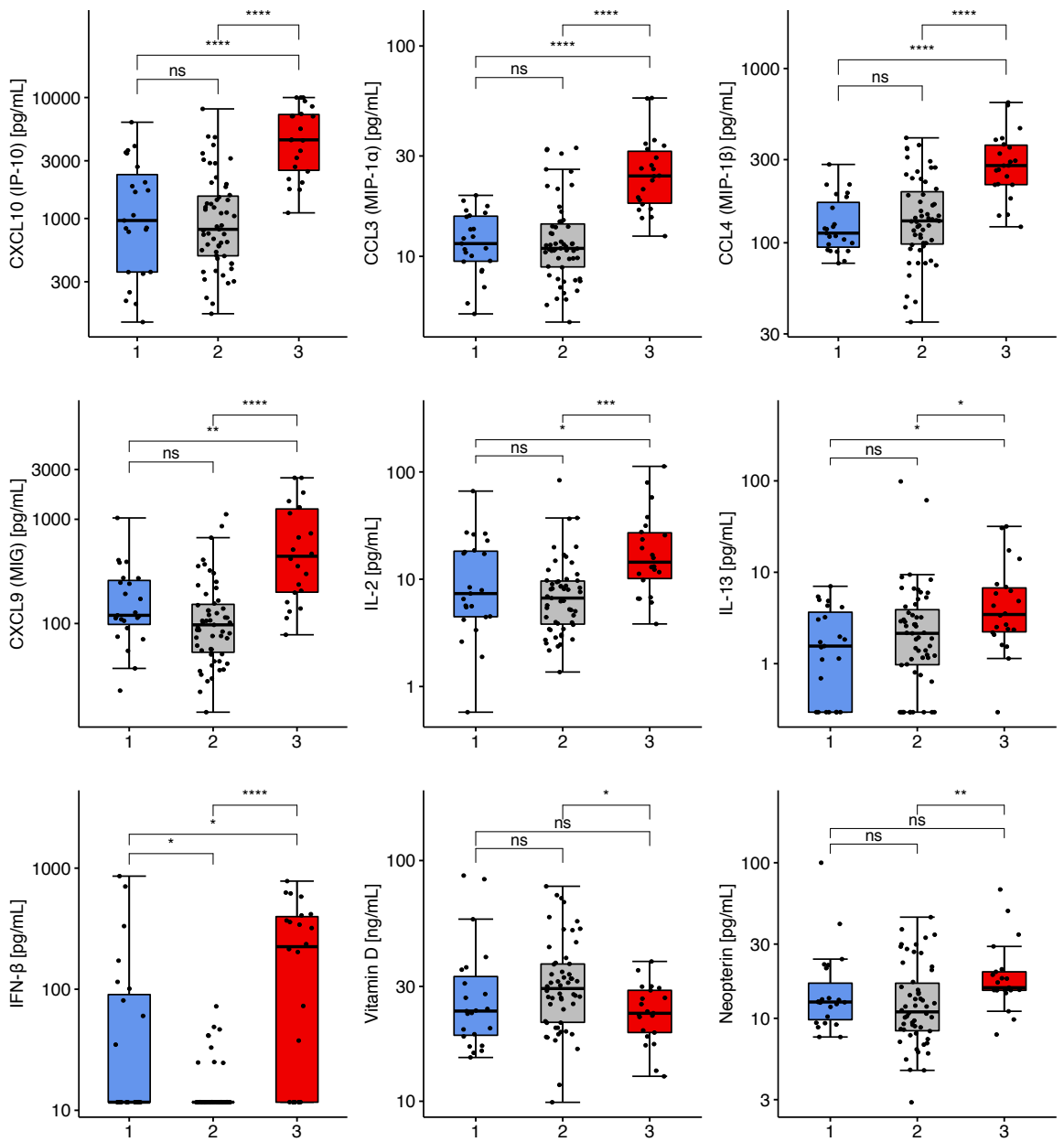
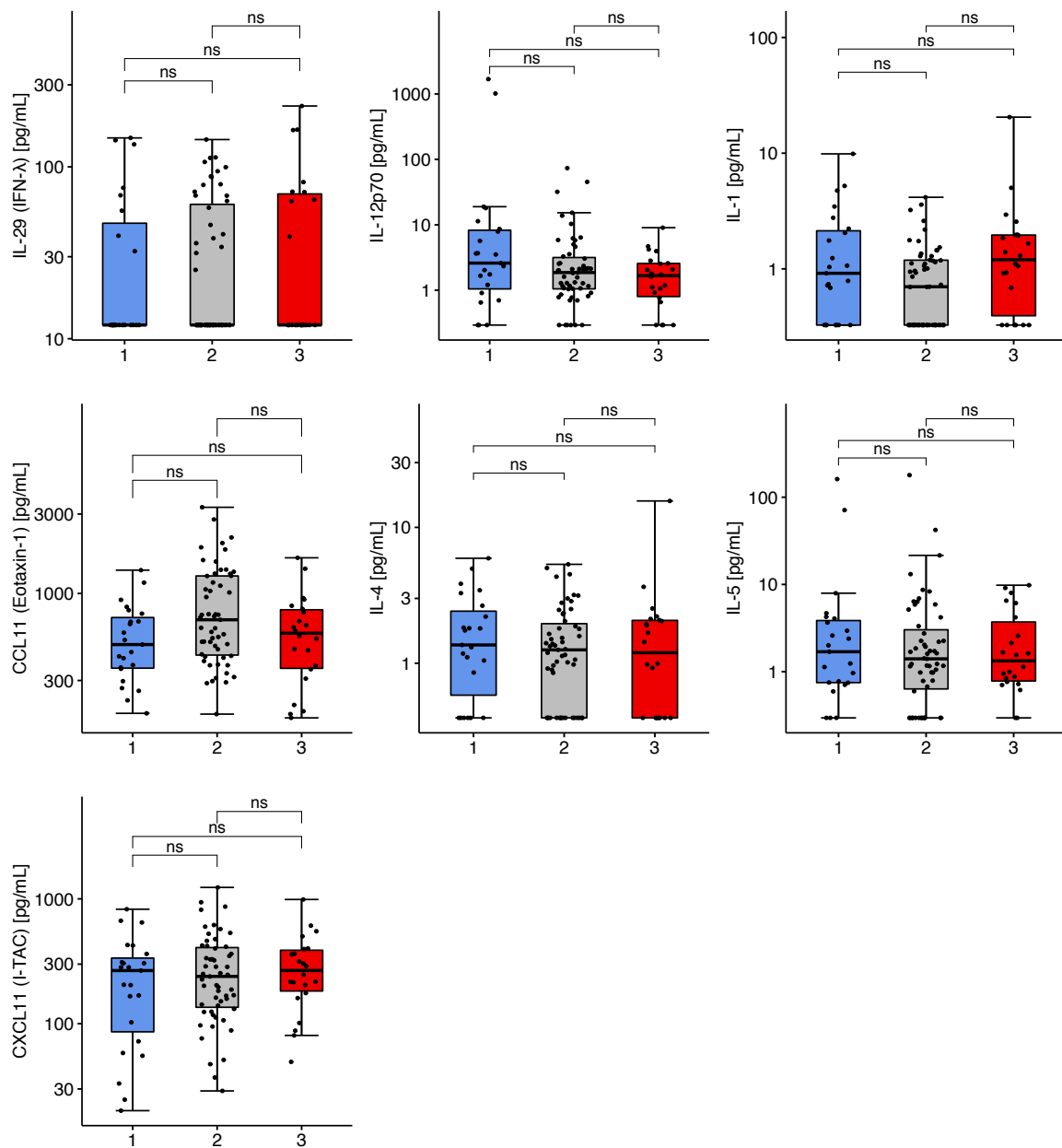


Figure D.2c



(d)

Fig. D.2 Boxplots showing distributions of cytokine and chemokine concentrations in each of the MOSAIC clusters. The horizontal axes labels and boxplot colours denote the clusters. All comparisons were made using Dunn's test with FDR correction of p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, ns non-significant

D.3 HARP-2

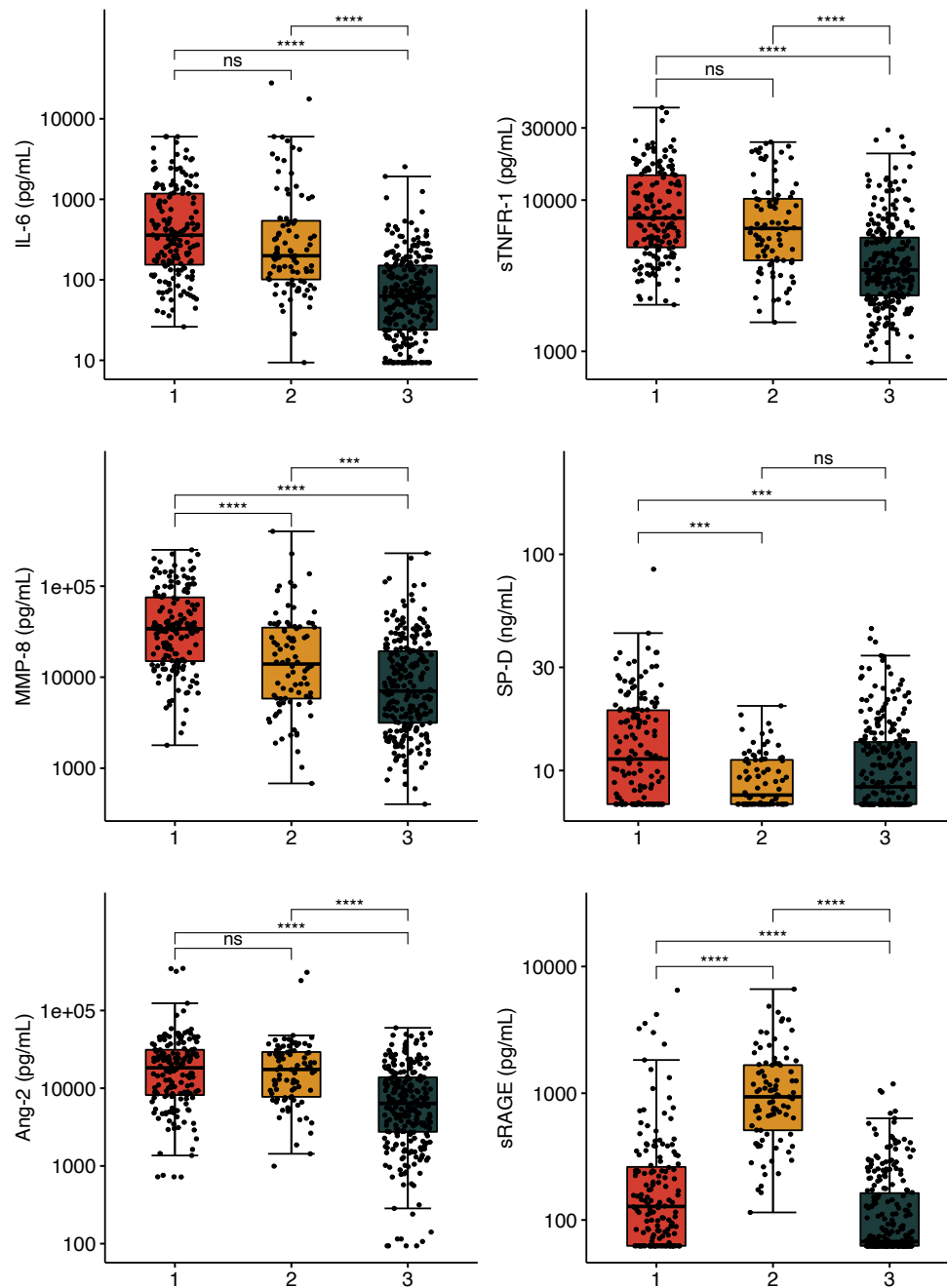


Fig. D.3 Boxplots showing distributions of protein biomarkers concentrations in patients from each of the HARP-2 clusters. The horizontal axes labels and boxplot colours denote the clusters. All comparisons were made using Dunn's test with FDR correction of p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, ns non-significant

E Fully labelled biplots

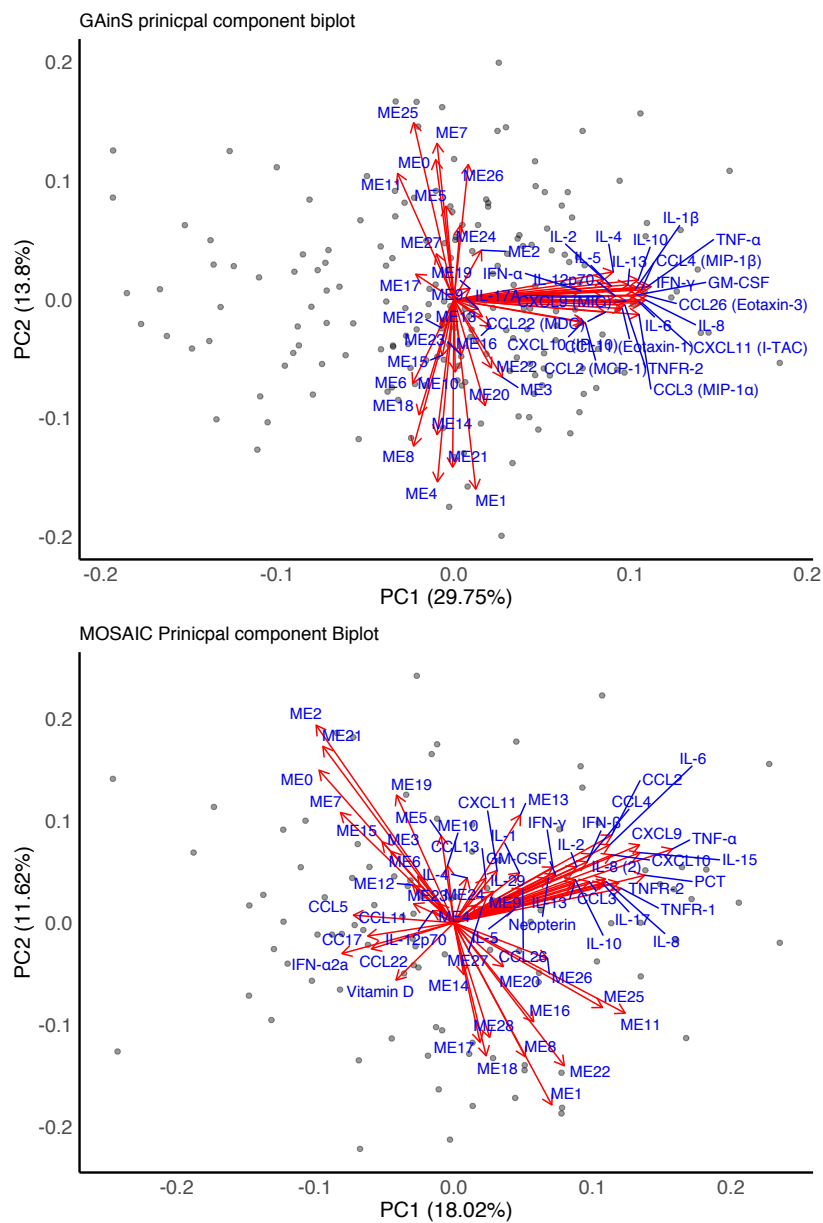


Fig. E.1 Biplot showing the directional loadings of individual module eigengenes (MEs) and cytokines following integration of these two data types. Red arrows indicate the direction of loadings for each variable.

F Ranked linear discriminators between clusters from the GAinS study

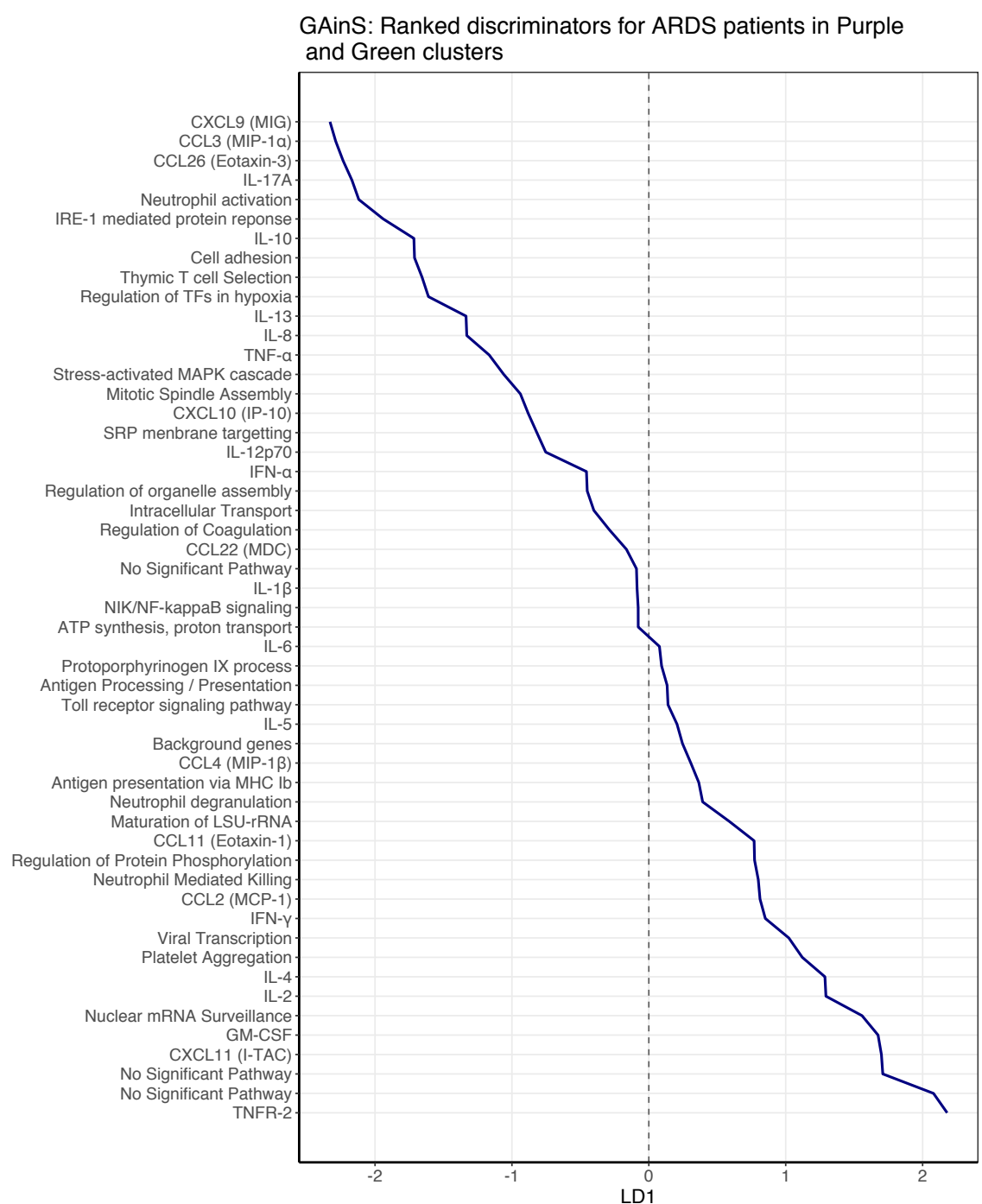


Fig. F.1 Line plot showing the full ranking of all discriminator variables between patients with ARDS in the ‘purple’ and ‘green’ GAINs clusters

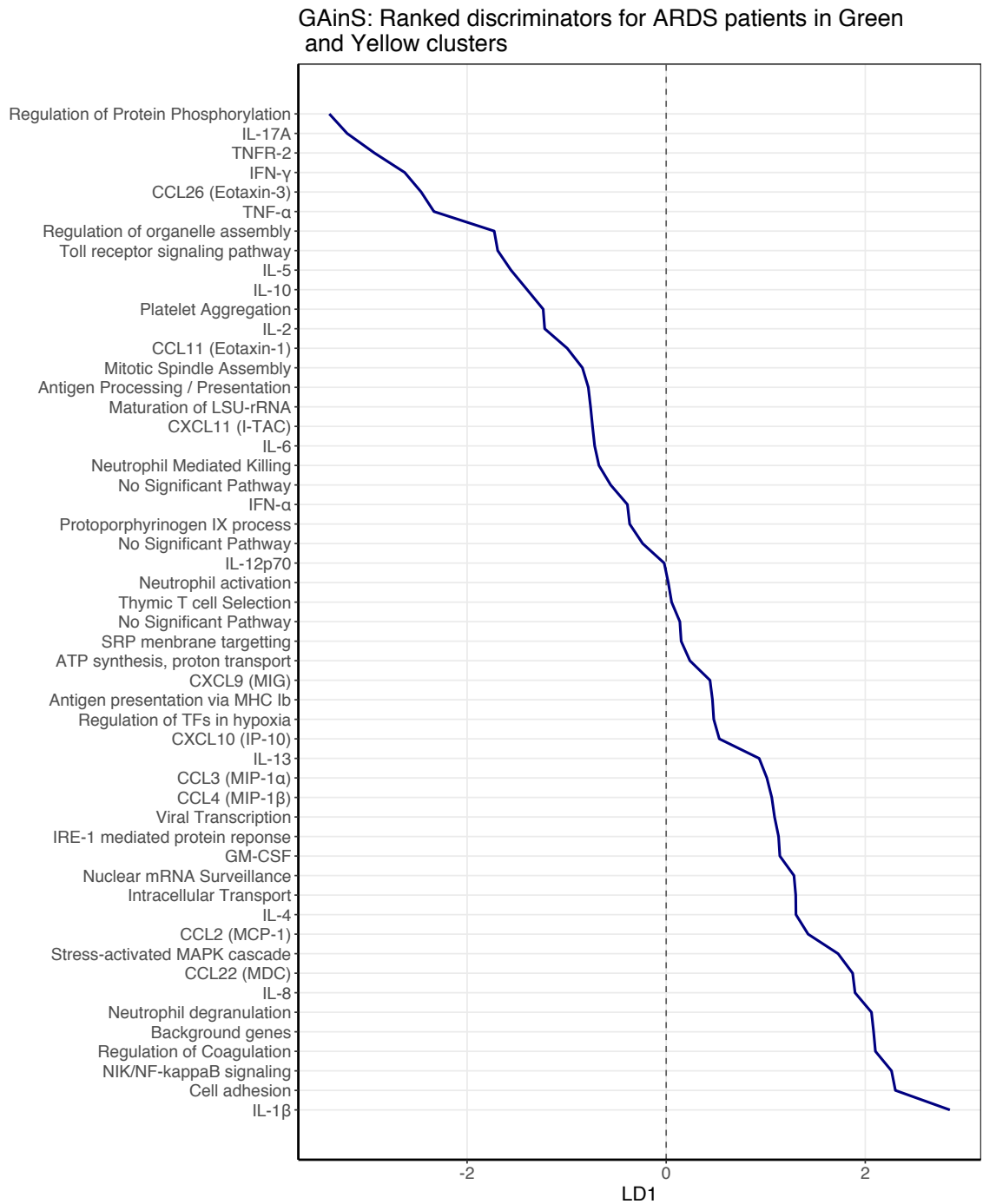


Fig. F.2 Line plot showing the full ranking of all discriminator variables between patients with ARDS in the 'green' and 'yellow' GAinS clusters

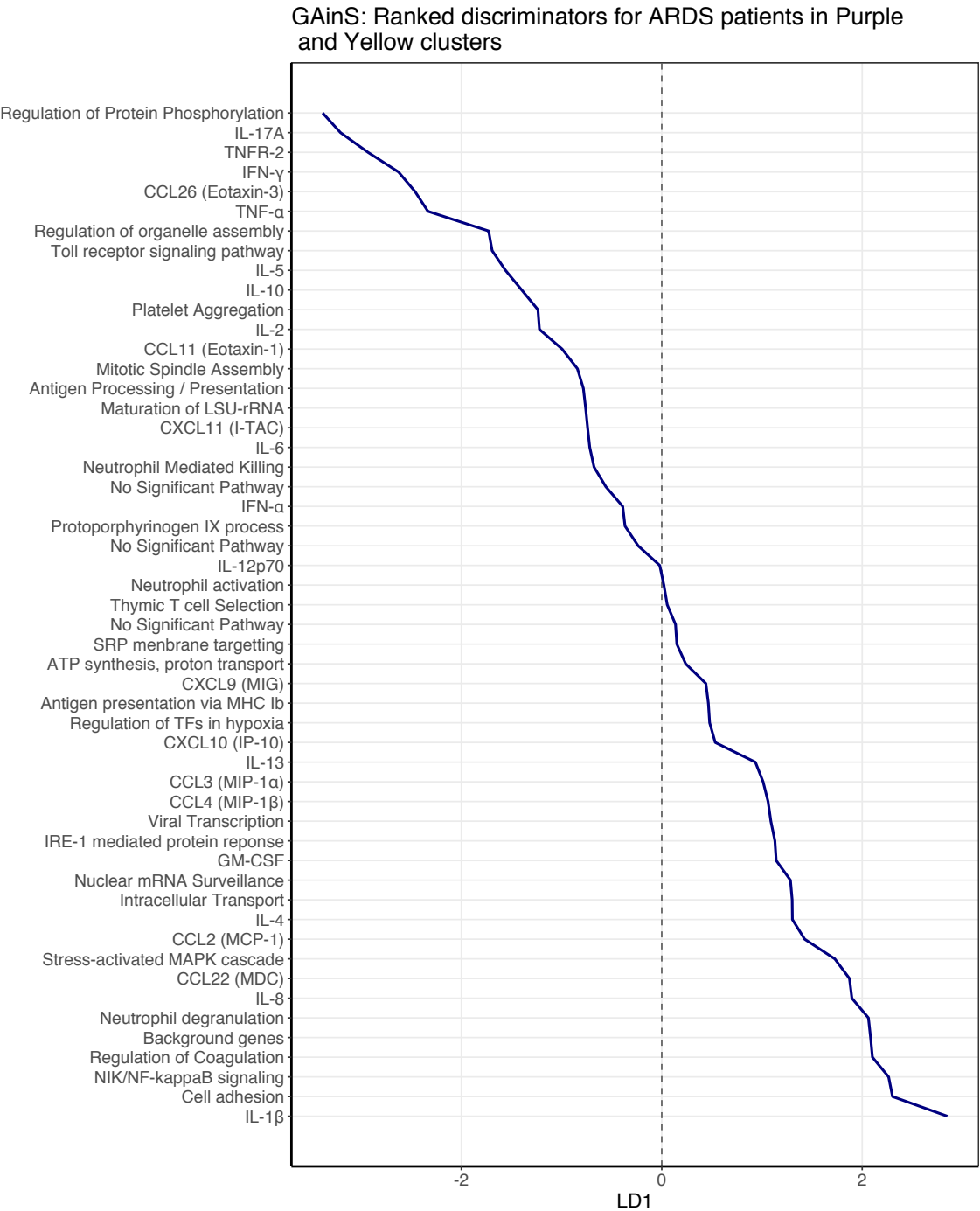


Fig. F.3 Line plot showing the full ranking of all discriminator variables between patients with ARDS in the 'purple' and 'yellow' GAINs clusters

G Ranked linear discriminators between clusters from the MOSAIC study

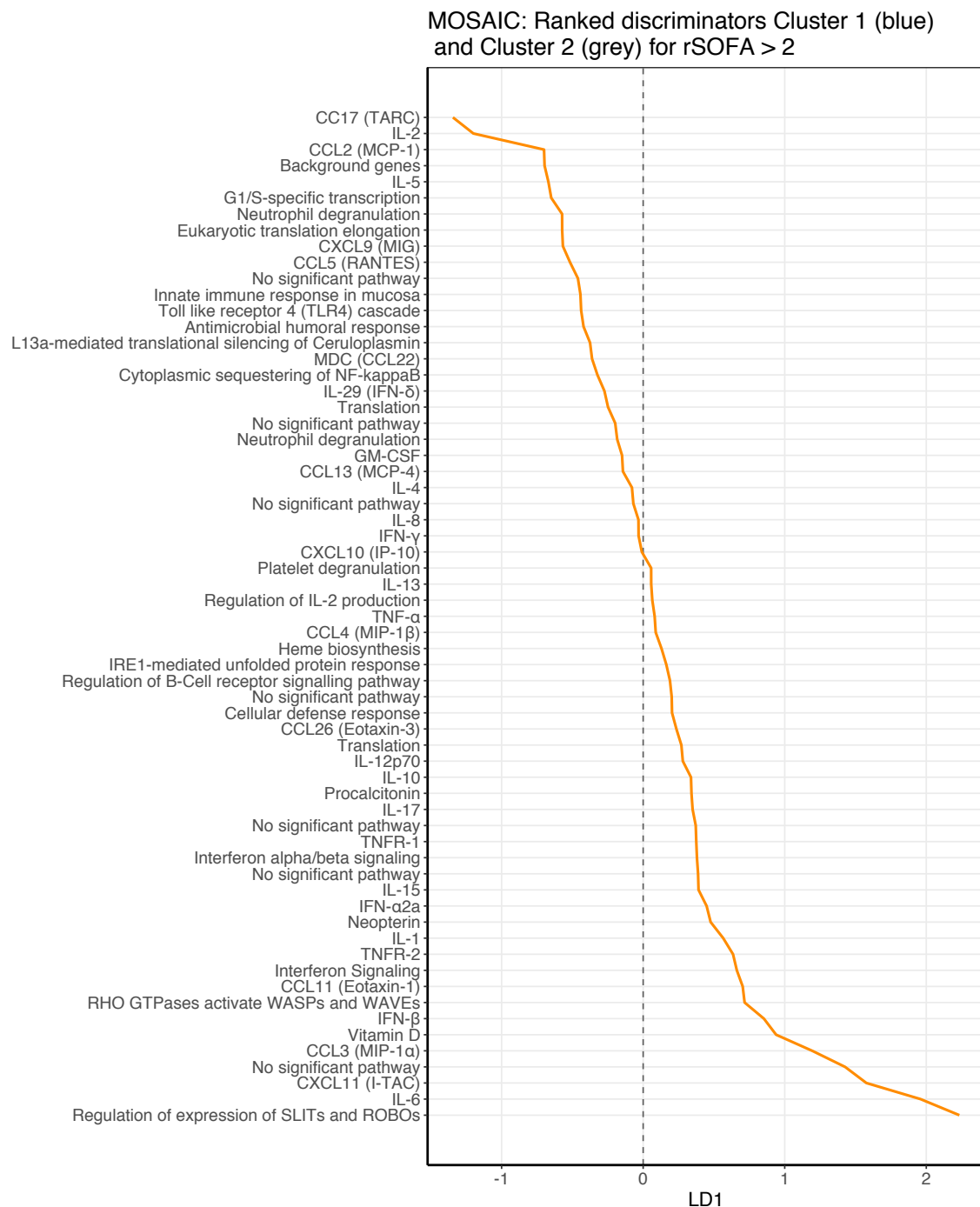


Fig. G.1 Line plot showing the full ranking of all discriminator variables between patients with rSOFA > 2 in the 'blue' and 'grey' MOSAIC clusters

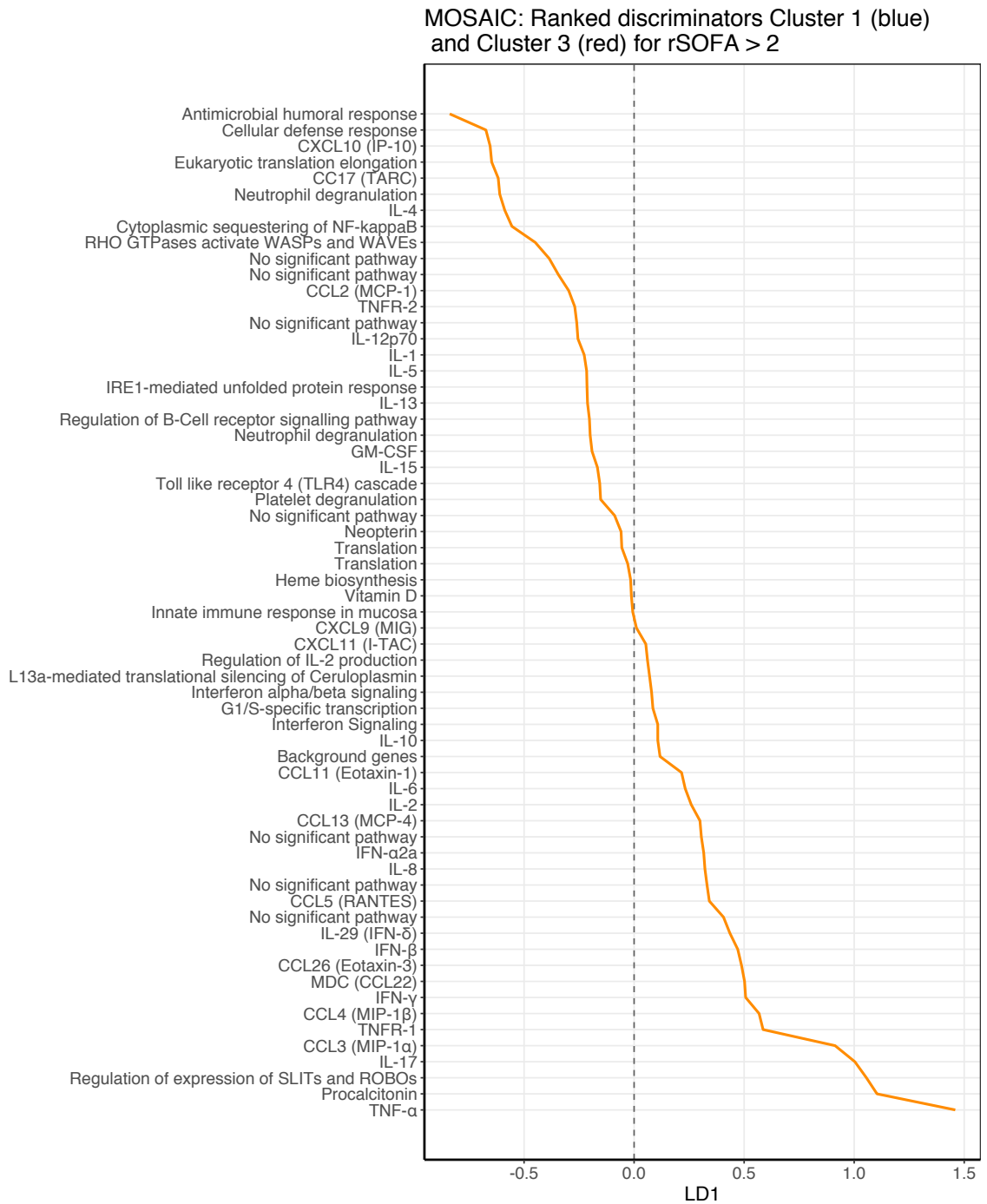


Fig. G.2 Line plot showing the full ranking of all discriminator variables between patients with rSOFA > 2 in the Blue and Red MOSAIC clusters

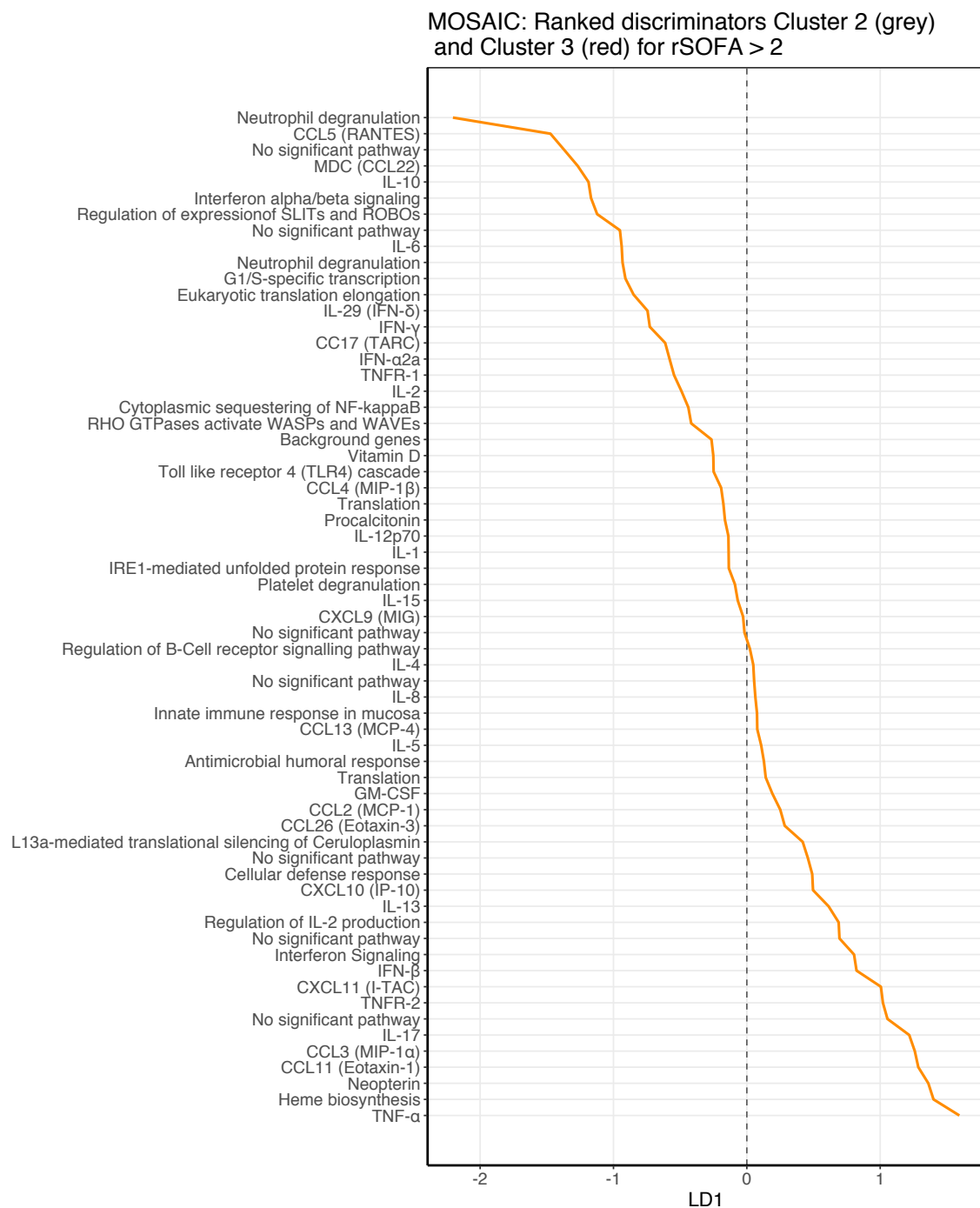


Fig. G.3 Line plot showing the full ranking of all discriminator variables between patients with rSOFA > 2 in the Red and Grey MOSAIC clusters

H Permissions

From: **ATS Permission Requests** permissions@thoracic.org
Subject: RE: Permission request for figure from Bos et al doi:10.1164/rccm.201809-1808OC
Date: 23 November 2020 at 15:48
To: Romit Samanta rs307@cam.ac.uk

Dear Dr. Samanta,

Thank you for your request. Because your request is for thesis use, permission for use of unmodified Figure 4 is granted at no charge. Please complete the below and use it beneath the material. Thank you.

Reprinted with permission of the American Thoracic Society.
Copyright © 2020 American Thoracic Society. All rights reserved.
Cite: Author(s)/Year/Title/Journal title/Volume/Pages.
The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society.

Thank you, and please let me know if you have any questions.

Best regards,

Megan

Megan Murphy
Production Coordinator
American Thoracic Society
25 Broadway, 4th Floor
New York, NY 10004
212-315-8643
mmurphy@thoracic.org

[Please donate](#) to the ATS to make COVID-19 resources accessible and support respiratory health professionals worldwide. Wishing you, your community, and all health care professionals safety during this difficult time.

From: Romit Samanta <rs307@cam.ac.uk>
Sent: Monday, November 23, 2020 8:09 AM
To: ATS Permission Requests <permissions@thoracic.org>
Subject: Permission request for figure from Bos et al doi:10.1164/rccm.201809-1808OC

Dear ATS Journals Permissions Team,

I am clinical PhD candidate at the University of Cambridge and wish to request permission to use the following figure in the introduction section for my thesis on "Endotype Discovery in Acute Respiratory Distress Syndrome (ARDS)" from this article by Bos *et al* published in AJRCCM in July 2019.

The figure will be reprinted in full, unedited, on a separate page with full acknowledgment as to its origin and permission for its reuse in this context. The appendix of the thesis will include details of communication with the journal.

The context of the figure is to demonstrate the relative biological heterogeneity of

Publisher authorisation for use of Figure 1.1

05/12/2020

Rightslink® by Copyright Clearance Center



RightsLink®



Home



Help



Email Support



Romit Samanta ▾

**The GAIT system: a gatekeeper of inflammatory gene expression**

Author: Rupak Mukhopadhyay, Jie Jia, Abul Arif, Partho Sarothi Ray, Paul L. Fox

Publication: Trends in Biochemical Sciences

Publisher: Elsevier

Date: July 2009

Copyright © 2009 Elsevier Ltd. All rights reserved.

Review OrderPlease review the order details and the associated [terms and conditions](#).

No royalties will be charged for this reuse request although you are required to obtain a license and comply with the license terms and conditions. To obtain the license, click the Accept button below.

Licensed Content

Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Biochemical Sciences
Licensed Content Title	The GAIT system: a gatekeeper of inflammatory gene expression
Licensed Content Author	Rupak Mukhopadhyay, Jie Jia, Abul Arif, Partho Sarothi Ray, Paul L. Fox
Licensed Content Date	July 2009
Licensed Content Volume	34
Licensed Content Issue	7
Licensed Content Pages	8
Journal Type	S&T

Order Details

Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No

About Your Work

Title	Endotype discovery in acute respiratory distress syndrome
Institution name	University of Cambridge
Expected presentation date	Jan 2021

Additional Data

Portions	Figure 2. The entire figure will be used, unedited with full attribution to its source.
----------	-----------------------------------------------------------------------------------------

Requestor Location

Requestor Location	Dr. Romit Samanta Box 157 Addenbrooke's Hospital Level 5, Department of Medicine Hills Road Cambridge, CB2 0QQ United Kingdom Attn: Dr. Romit Samanta
--------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Tax Details

Publisher Tax ID	GB 494 6272 12
------------------	----------------


\$ Price

Total	0.00 GBP
-------	----------

https://s100.copyright.com/AppDispatchServlet

1/2

Publisher authorisation for use of Figure 4.6




Home

Help

Email Support

Romit Samanta

RightsLink®



Targeting Robo4-Dependent Slit Signaling to Survive the Cytokine Storm in Sepsis and Influenza

Author:
Niyall R. London,WeiQuan Zhu,Fernando A. Bozza,Matthew C. P. Smith,Daniel M. Greif,Lise K. Sorensen,Luming Chen,Yuuki Kaminoh,Aubrey C. Chan,Samuel F. Passi,Craig W. Day,Dale L. Barnard,Guy A. Zimmerman,Mark A. Krasnow,Dean Y. Li

Publication: Science Translational Medicine

Publisher: The American Association for the Advancement of Science

Date: Mar 17, 2010

Copyright © 2010, Copyright © 2010, American Association for the Advancement of Science

Order Completed

Thank you for your order.

This Agreement between Dr. Romit Samanta ("You") and The American Association for the Advancement of Science ("The American Association for the Advancement of Science") consists of your license details and the terms and conditions provided by The American Association for the Advancement of Science and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number4962990408577

License dateDec 06, 2020

Licensed Content

Licensed Content PublisherThe American Association for the Advancement of Science

Licensed Content PublicationScience Translational Medicine

Licensed Content TitleTargeting Robo4-Dependent Slit Signaling to Survive the Cytokine Storm in Sepsis and Influenza

Licensed Content AuthorNiyall R. London,WeiQuan Zhu,Fernando A. Bozza,Matthew C. P. Smith,Daniel M. Greif,Lise K. Sorensen,Luming Chen,Yuuki Kaminoh,Aubrey C. Chan,Samuel F. Passi,Craig W. Day,Dale L. Barnard,Guy A. Zimmerman,Mark A. Krasnow,Dean Y. Li

Licensed Content DateMar 17, 2010

Licensed Content Volume2

Licensed Content Issue23

Order Details

Type of UseThesis / Dissertation

Requestor typeScientist/individual at a research institution

FormatPrint and electronic

PortionFigure

Number of figures/tables1

About Your Work

TitleEndotype discovery in acue respiratory distress syndrome

Institution nameUniversity of Cambridge

Expected presentation dateJan 2021

Additional Data

PortionsFigure 6.

Requestor Location

Requestor LocationDr. Romit Samanta
Box 157 Addenbrooke's Hospital
Level 5, Department of Medicine
Hills Road
Cambridge, CB2 0QQ
United Kingdom
Attn: Dr. Romit Samanta

Tax Details

Price

Total0.00 GBP

Total: 0.00 GBP

CLOSE WINDOW

ORDER MORE

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions

Comments? We would like to hear from you. E-mail us at customer@copyright.com

Publisher authorisation for use of Figure 4.19